UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

APERFEIÇOAMENTO DO ALGORITMO COLÔNIA DE FORMIGAS PARA O DESENVOLVIMENTO DE MODELOS QUIMIOMÉTRICOS

DISSERTAÇÃO DE MESTRADO

Carolina de Marco Pessoa

Porto Alegre

2015

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

APERFEIÇOAMENTO DO ALGORITMO COLÔNIA DE FORMIGAS PARA O DESENVOLVIMENTO DE MODELOS QUIMIOMÉTRICOS

Carolina de Marco Pessoa

Dissertação de Mestrado apresentada como requisito parcial para obtenção do título de Mestre em Engenharia

Área de concentração: Pesquisa e Desenvolvimento de Processos

Linha de Pesquisa: Projeto, Simulação, Modelagem, Controle e Otimização de Processo.

Orientador:
Prof. Dr. Jorge Otávio Trierweiler

Porto Alegre

2015

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

A Comissão Examinadora, abaixo assinada, aprova a Dissertação *Aperfeiçoamento do Algoritmo Colônia de Formigas para o Desenvolvimento de Modelos Quimiométricos*, elaborada por Carolina de Marco Pessoa, como requisito parcial para obtenção do Grau de Mestre em Engenharia.

Comissão Examinadora:

Prof. Dr. Marcelo Farenzena

Prof. Dr. Marco Flôres Ferrão

Prof. Dr. Michel Jose Anzanello

Resumo

O desenvolvimento e aperfeiçoamento de métodos de otimização são pontos de profundo interesse em todas as áreas de pesquisa. Tais técnicas muitas vezes envolvem a aquisição de métodos de controle novos ou melhores, o que está diretamente ligado a duas tarefas importantes: a escolha de formas eficientes de monitoramento do processo e a obtenção de modelos confiáveis para a variável de interesse a partir de dados experimentais. Graças às suas diversas vantagens, os sensores óticos vêm sendo amplamente aplicados na primeira tarefa. Uma vez que é possível a utilização de vários tipos de espectroscopia através deste tipo de sensor, modelos capazes de lidar com dados espectrais estão se tornando cada vez mais atraentes. A segunda tarefa, por sua vez, depende não só de quais preditores são utilizados na construção do modelo, mas também de quantos. Como a qualidade do modelo depende também do número de variáveis selecionadas, é importante desenvolver métodos que identifiquem aqueles que explicam o máximo possível da variabilidade dos dados. O método de otimização Colônia de Formigas (ACO) aparece como uma ferramenta bastante útil na seleção de variáveis, podendo-se encontrar muitas variações desse algoritmo na literatura. O propósito deste trabalho é desenvolver métodos de seleção de variáveis com base no algoritmo ACO, conceitos estatísticos e testes de hipóteses. Para isso, diversos critérios de decisão foram implementados nas etapas do algoritmo referentes à atualização de trilha de feromônios (C1) e à seleção de modelos (C2). A fim de estudar estas modificações, foram realizados dois estudos de caso: o primeiro na área de bioprocessos e o segundo na área de caracterização de alimentos. Ambos os estudos mostraram que, em geral, os modelos com menores erros são obtidos utilizando-se métricas dos componentes do modelo, tal como o tamanho do intervalo de confiança de cada parâmetro e o teste-t de hipóteses. Além disso, a modificação do critério de seleção de modelos parece não interferir significativamente no resultado final do algoritmo. Por último, foi feito um estudo da aplicação dessas versões do ACO no campo de caracterização de combustíveis, mais especificamente diesel, associando-se duas análises espectroscópicas para predição do conteúdo de enxofre. Algumas das versões desenvolvidas mostraram-se superior ao algoritmo ACO utilizado como base para este trabalho, proposto por Ranzan (2014), e todas os versões forneceram melhores resultados na quantificação de enxofre que aqueles obtidos por PCR. Dessa forma, comprova-se a potencialidade de métricas implementadas no algoritmo ACO, associadas à espectroscopia, na seleção de preditores significativos.

Palavras-chave: otimização colônia de formigas, espectroscopia de fluorescência 2D, espectroscopia NIR, diesel.

Abstract

The development and improvement of optimization methods are points of deep interest in all areas of research. These techniques are often related to the acquisition of new or better control methods, which are directly attached to two importante tasks: choosing efficient forms of process monitoring and obtaining reliable models for the monitored variable from experimental data. Due to their several advantagens, optical sensors are being widely applied in the first task. Since several types of spectroscopy are possible through this type of sensor, models capable of dealing with spectral data are becoming increasingly attractive. The second task depends not only on which predictors are used in the model, but also on how many. Since the quality of the model depends on the number of selected variables, it is important to develop methods that identify those that explain the greater amount of data variability as possible, without compromising the reliability of the model. The Ant Colony Optimization is an important tool for variable selection, being possible to find a lot of variations of this method in literature. The purpose of this work is to develop a method of variable selection based on the Ant Colony Optimization (ACO) algorithm, statistical concepts and hypothesis testing. For this purpose, several decision criteria for trail update (C1) and model selection (C2) were implemented within the routine. In order to study these modifications, two case study was conducted: one related to bioprocess monitoring and another one envolving the characterization of food products. Both studies showed that, in general, the models with the lowest errors were obtained through the use of model component metrics, such as the length of the confidence interval associated with each parameter and the t hypothesis test. Besides, the modification of the model selection criterion doesn't seem to affect the algorithm final result. Finally, the aplicattion of these methods in the field of fuels characterization, specifically diesel fuel, was studied, associating two spectroscopical analyses in order to predict the sulfur content. Some of the new developed methods appeared to be better than the ACO algorithm used as basis in this work, proposed by Ranzan (2014), and all methods showed better results than those from the models constructed by PCR. Thus, it is proved the high potencial of using different metrics within ACO algorithm, associated with spectroscopy, in order to select significative predictors.

Key-words: ant colony optimization, 2D fluorescence spectroscopy, NIR spectroscopy, diesel.

“The only true wisdom is in knowing you know nothing.”

- Socrates

Agradecimentos

Primeiramente, agradeço aos meus pais por todo carinho e apoio (financeiro e emocional) ao longo desses anos, bem como por todo suporte psicológico nos momentos de crise. Sou grata ao meu irmão Felipe pelo companheirismo, inspiração e pelas críticas construtivas. Agradeço todos os dias por tê-los como família, obrigada por tudo.

Agradeço também ao meu professor orientador Dr. Jorge Trierweiler, por todos os conselhos, sugestões e ajuda prestada ao longo da preparação deste trabalho.

Ao colega e agora professor Dr. Cassiano Ranzan, essencial para a realização deste trabalho, pelas incontáveis vezes que me prestou ajuda e pela boa vontade com que o fez.

Ao meu amado companheiro Renan, por quem tenho profunda admiração, por todo carinho, humor e compreensão, bem como por todo conhecimento teórico compartilhado que me pertiu concluir este trabalho. Sou grata à pós-graduação por ter me proporcionado a oportunidade de conhecer alguém tão incrível e hoje tão importante na minha vida.

Aos meus colegas do departamento, pela amizade, pelo auxílio prestado no dia-a-dia e pelos momentos de descontração, fazendo meus dias mais divertidos.

Finalmente, à Mint e ao Tintim pelo amor incondicional e por preencher meus dias com muita alegria, tornando a vida mais leve.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIAÇÕES

ACO	Otimização Colônia de Formigas (*Ant Colony Optimization*)
NIR	Infravermelho Próximo (*Near Infrared*)
MIR	Infravermelho Médio (*Middle Infrared*)
MLR	Regressão Multilinear (Multilinear Regression)
PC	Componente Principal (*Principal Component*)
PCA	Análise por Componentes Principais (*Principal Component Analysis*)
PCR	Regressão por Componentes Principais (*Principal Component Regression*)
PLS	Mínimos Quadráticos Parciais (*Partial Least Squares*)
PLSR	Regressão com Mínimos Quadráticos Parciais (*Partial Least Squares Regression*)
PSCM	Modelagem com Componentes Espectrais Puros *(Pure Spectral Component Modeling*)
R^2	Coeficiente de determinação
R_a^2	Coeficiente de determinação ajustado
RMSEP	Raiz Quadrada do Erro Quadrático Médio da Predição (*Root Mean Square Error of Prediction*)
RR	Coeficiente de determinação modificado
SNV	Variação Normal Padrão (*Standard Normal Variate*)
SSE	Soma dos Erros Quadráticos (*Sum of Squared Errors*)
SSR	Soma dos Quadrados de Regressão (*Regression Sum of Squares*)
SST	Soma dos Quadrados Totais (*Total Sum of Squares*)

NOTAÇÃO E SIMBOLOGIA

ρ_{Fi}	Densidade de feromônio relativa do elemento espectral *i*
F_i	Quantidade de feromônio associada ao elemento espectral *i*
C1	Critério de atualização da trilha de feromônios no algoritmo ACO
C2	Critério de seleção de modelos no algoritmo ACO
N	Número de elementos espectrais na matriz de entrada
C_{Fi}	Densidade de feromônios acumulada
$\hat{y}_i$	Valor predito para a variável de saída
y_i	Valor medido para a variável de saída
n	Número de amostras utilizadas na calibração
e_i	Diferença entre valor medido e valor predito pelo modelo
$\bar{y}$	Média dos valores medidos para a variável de saída
k	Tamanho do modelo
β_j	Coeficiente associado ao componente *j*
t	Valor da estatística t para teste de hipótese com distribuição t-student
$t_{\frac{\alpha}{2}}$	Estatística t associada a um grau de confiança $\frac{\alpha}{2}$
b_j	Estimador do parâmetro β_j
c_{jj}	Variância de b_j
LCI	Tamanho do intervalo de confiaça (*Length of Confidence Interval*)
F	Valor da estatística F para teste de hipótese com distribuição normal
k_{sub}	Tamanho do submodelo

Capítulo 1 – Introdução

O presente capítulo apresenta uma introdução ao assunto de que trata esta dissertação, mediante o estabelecimento dos aspectos que motivaram este trabalho, dos objetivos de sua realização e da forma como o mesmo está estruturado.

1.1 Motivação

Frente à crescente demanda por produtos industrializados, empresas de diversos segmentos têm procurado formas de melhoria dos processos a fim de aumentar a competitividade de seus produtos. Diante disso, o estudo e o aperfeiçoamento de técnicas de otimização são relevantes para todas as áreas de atuação, uma vez que permitem reduções nos custos de produção e um aumento na conversão das reações, além de garantirem a qualidade do produto desejado. Tais técnicas estão frequentemente relacionadas ao aprimoramento do controle do processo que, por sua vez, está ligado a duas grandes frentes de atuação: a obtenção de um bom modelo para a variável de interesse e a escolha de técnicas eficientes de monitoramento do processo (Yamuna e Ramachandra, 1999).

Basicamente, modelagem consiste em encontrar uma relação causal entre as variáveis de um processo ou fenômeno. A análise de regressão, combinada com técnicas estatísticas para quantificar a confiança do modelo, aparece como o principal instrumento utilizado para este fim (Sykes, 1993). Entre as diferentes formas de regressão, a linear é certamente a mais utilizada devido à sua simplicidade. Neste contexto, técnicas como Mínimos Quadrados Parciais (PLS) e Regressão por Componentes Principais (PCR) aparecem como dois dos principais métodos de regressão linear, úteis na análise quantitativa de dados (Geladi *et al.*, 2004).

No entanto, a qualidade do modelo depende não só das variáveis utilizadas na regressão, mas também da quantidade desses preditores: um pequeno subconjunto de variáveis de previsão é muitas vezes preferível frente ao uso de todos os dados disponíveis. Isso reduz os custos e o tempo gasto nas medições, tende a apresentar modelos com uma interpretação física mais simples e, no caso de regressão linear múltipla (MLR), reduz a incerteza de predição, uma vez que esta incerteza aumenta com a

razão entre o número de variáveis explanatórias e o número de amostras usadas na calibração (Brown, Tauler e Walczak, 2009).

O monitoramento do processo, por sua vez, é altamente dependente do tipo e da qualidade dos sensores utilizados. Uma excelente alternativa para essa tarefa requer que cada parâmetro monitorado em um sistema de controle possa ser obtido através do emprego de um pequeno sensor eletrônico in situ, sendo que esse sensor deve ser capaz de satisfazer certos requisitos como (Lindemann *et al.*, 1998; Skibsted *et al.*, 2001; Wong *et al.*, 2014):

- Ser inerte ao processo;
- Ser passível de esterilização no local ou autoclavagem;
- Permitir limpeza no local, além de ser facilmente lavado manualmente;
- Possuir sensibilidade e resolução suficientes, necessárias ao sistema de monitoramento e controle;
- Ser facilmente calibrável através da utilização de soluções e procedimentos padrões aplicados em laboratórios;
- Apresentar pouco ruído e baixa interferência causada por fatores externos;
- Ser robusto para operar nas condições de processos;
- Apresentar alta relação benefício/custo.

Com base nesses atributos, vem surgindo um crescente interesse no desenvolvimento e aplicação de sensores ópticos na caracterização de processos, especialmente no segmento biotecnológico. A atratividade de sensores ópticos em relação a sensores convencionais reside nas vantagens de sua utilização, devidas principalmente a sua imunidade eletromagnética às interferências, permitindo a construção de equipamentos de monitoramento remoto, contínuos, com tamanho reduzido e capazes de mensurar simultaneamente diversas grandezas de interesse (Skibsted *et al.*, 2001; Wong *et al.*, 2014). Além disso, uma vez que os sensores ópticos podem ser conectados através de janelas de vidro nos reatores, esse é um método in situ e não invasivo, que fornece medições em tempo real (Scheper *et al.*, 1999; Hantelmann *et al.*, 2006).

O interesse em espectrometria é também elevado no ramo de combustíveis, devido ao seu grande potencial de inferência de diversas propriedades físico-químicas de forma rápida, eficiente e não destrutiva, conforme mostrado por diversos autores (Santos Jr *et al.*, 2005; Felizardo *et al.*, 2007; Baptista *et al.*, 2008; Balabin, Lomakina e Safieva, 2011; Silva *et al.*, 2015). A utilização dessas técnicas pode então garantir que propriedades como densidade, viscosidade, conteúdo de enxofre, composição das blendas, lubricidade, ponto de fulgor, entre outras, estejam em conformidade com a legislação vigente para aquele combustível.

O uso de sensores ópticos pode estar associado a vários tipos de espectroscopia, fato que torna tão atraentes os modelos capazes de lidar com dados espectrais. Dentre os métodos espectroscópicos mais utilizados industrialmente pode-se citar a Espectroscopia no Infravermelho Médio e Próximo (MIR/NIR), a Espectroscopia RAMAN e a Espectroscopia de Fluorescência 2D. Neste contexto, o desenvolvimento de um método capaz de trabalhar com dados espectrais de modo a identificar os componentes ou regiões espectrais melhor relacionadas com a variável de resposta pode permitir o desenvolvimento de sensores ópticos adaptados a processos específicos, melhorando seu controle e, consequentemente, tornando-os mais eficientes e econômicos.

Assim, o sucesso das estratégias de controle e, portanto, do processo, recai fortemente sobre as variáveis selecionadas para a construção do modelo empregado. Nesse aspecto, diversas técnicas quimiométricas têm sido propostas como meios de identificar os melhores preditores, dentre as quais se pode citar os trabalhos de Ranzan *et al.* (2014), Goodarzi e Dos Santos Coelho (2014), Hapfelmeier e Ulm (2014), Zhang *et al.* (2014) e Araki *et al.* (2015).

Uma abordagem bastante comum na seleção de variáveis é a combinação de critérios adequados que avaliam a qualidade de um subconjunto de preditores combinados com um algoritmo que optimiza estes critérios (Brown, Tauler e Walczak, 2009). Tal abordagem é utilizada neste trabalho, aplicando-se o algoritmo Colônia de Formigas (Ant Colony Optimization - ACO) como método de otimização.

O algoritmo ACO é baseado no comportamento coletivo hipotético de formigas quando em busca de fontes de alimento. Nessa abstração, as formigas secretam feromônios que marcam os caminhos percorridos por elas, mas que evaporam ao longo do tempo. As formigas que viajam pelo caminho mais curto retornam à colônia mais rapidamente, de modo que o trajeto percorrido por estes indivíduos tem uma maior concentração de feromônio. Essa trilha funciona como um chamariz para outras formigas e, com o tempo, todos os indivíduos da colônia tendem a utilizar este caminho ótimo (mais curto) (Allegrini & Olivieri 2011). Esse algoritmo é composto por duas etapas principais, uma relacionada à atualização da trilha de feromônios e outra relacionada à comparação dos modelos gerados (Ranzan, C., 2014).

1.2 Objetivos do trabalho

O presente trabalho tem como objetivo principal a proposta de uma nova versão do algoritmo de otimização ACO, modificado com base em conceitos estatísticos e testes de hipóteses. Assim, são introduzidas no algoritmo novas métricas de avaliação em duas das suas principais etapas: a atualização da trilha de feromônios e a seleção de modelos. Busca-se, portanto, avaliar uma série de questões a respeito destas modificações, a saber:

1) Determinar se a alteração da métrica de comparação e seleção dos modelos no algoritmo modifica os resultados alcançados;

2) Verificar se a utilização de métricas de avaliação de preditores introduz melhorias no algoritmo frente ao uso de métricas do modelo total;

3) Verificar se existe uma combinação ótima de métricas dessas duas etapas, bem como se tal combinação depende do conjunto de dados espectroscópicos utilizados;

4) Determinar se os resultados encontrados pelos algoritmos modificados são superiores àqueles obtidos pelo algoritmo original, proposto por Ranzan, C. (2014);

5) Determinar qual combinação de critérios fornecem os melhores resultados quando aplicados à determinação de propriedades de combustíveis, especificamente conteúdo de enxofre em diesel;

6) Verificar se a aplicação dessas novas versões do algoritmo à combinação de dados provenientes de técnicas espectrométricas diferentes introduz melhorias nos resultados.

A fim de atingir os quatro primeiros objetivos, foram realizados dois estudos de caso, o primeiro visando à predição de concentração de biomassa em um meio fermentativo utilizando-se espectroscopia de fluorescência bidimensional e o segundo à predição de conteúdo protéico em amostras de farinha por meio de espectroscopia de infravermelho próximo (NIR). As últimas duas questões são esclarecidas pelo capítulo 5 desta dissertação, que trata da combinação de técnicas espectrométricas para predição do conteúdo de enxofre em combustível diesel.

1.3 Estrutura da Dissertação

Esta dissertação contém cinco capítulos além do atual capítulo referente à introdução, totalizando seis capítulos.

Inicialmente, uma revisão bibliográfica é apresentada no capítulo 2 abordando os principais conceitos em que este trabalho se baseia, tais como métodos espectroscópicos, modelos quimiométricos, métricas de qualidade de modelo e caracterização de combustível diesel. Ao longo deste capítulo são também apresentadas como e em que áreas os diferentes trabalhos existentes na literatura têm aplicado esses conceitos.

O capítulo 3 descreve as modificações realizadas no algoritmo ACO que será aplicado nos estudos de caso. Este capítulo apresenta também a legenda utilizada para interpretação e discussão dos resultados.

O capítulo 4 trata dos estudos de caso utilizados para testar as modificações abordadas anteriormente. Este capítulo subdivide-se em duas seções, uma referente ao estudo de caso utilizando espectroscopia de fluorescência para predição de concentração de biomassa em um meio fermentativo e outro referente ao caso utilizando espectroscopia NIR para predição de conteúdo protéico em farinha. Cada seção apresenta a descrição do experimento, abordando materiais e métodos utilizados no respectivo estudo, e a discussão dos resultados encontrados para aquele conjunto de dados.

No capítulo 5, o algoritmo ACO modificado apresentado no capítulo 3 e estudado no capítulo 4 é aplicado na determinação do conteúdo de enxofre em amostras de combustível diesel, utilizando-se a espectroscopia de fluorescência bidimensional associada à espectroscopia no infravermelho próximo.

O capítulo 6 apresenta as conclusões finais e sugestões de trabalhos futuros.

Capítulo 2 – Revisão Bibliográfica

A fim de proporcionar um melhor entendimento do presente trabalho, este capítulo apresenta uma revisão dos conceitos fundamentais utilizados na sua elaboração. Inicialmente são discutidas as principais técnicas espectrométricas utilizadas para aquisição de dados, seguido de um subcapítulo referente aos modelos quimiométricos utilizados para a construção de modelos a partir desses dados espectrais. Em seguida, as diferentes métricas usadas para quantificação da qualidade desses modelos são discutidas e, por último, é feita uma breve revisão da caracterização de combustível diesel, relacionada ao capítulo 5 desta dissertação.

2.1 Técnicas espectrométricas

A maioria dos métodos analíticos rápidos baseados nas propriedades físicas dos materiais são métodos espectrométricos, baseados na espectroscopia atômica e molecular. No entanto, a espectroscopia pode ser dividida em dois grandes grupos: espectroscopia fotônica, que se baseia no estudo da interação de ondas eletromagnéticas com a matéria, e espectroscopia de partículas, representada por espectroscopia de massa e métodos derivados. O primeiro grupo abrange métodos com grande potencial de utilização em controle rápido de processos e, portanto, é o que será abordado na presente seção (Skoog, Holler e Crouch, 2007).

As técnicas espectrométricas estão intrinsicamente relacionadas com o conceito de energia. Para átomos ou íons na forma elementar, a energia de qualquer estado provém da movimentação de elétrons ao redor do núcleo positivamente carregado. Porém, além desses estados eletrônicos, as moléculas também possuem estados associados à energia das vibrações interatômicas e da rotação molecular em torno de seu centro de massa (Skoog, Holler e Crouch, 2007).

Em análises espectroscópicas, a amostra é estimulada aplicando-se energia na forma de calor, energia elétrica, luz, partículas ou por uma reação química. Previamente ao

estímulo, o analito se encontra predominantemente em seu estado fundamental, ou seja, de menor energia. O estímulo então faz com que algumas espécies do analito sofram uma transição para um estado excitado de energia. Informações acerca do analito são obtidas através da análise da radiação emitida quando se retorna ao estado fundamental ou a quantidade de radiação eletromagnética absorvida ou espalhada decorrente da excitação (Skoog, Holler e Crouch, 2007).

Os métodos espectrométricos mais utilizados estão, portanto, baseados na radiação eletromagnética, podendo ser classificados de acordo com a energia envolvida na medição. A radiação eletromagnética, da qual a luz visível é apenas uma pequena parte, ocorre na forma de ondas que se propagam a partir de uma fonte e que se movem em linha reta quando não refletidas ou refratadas (Sun, 2009). Dessa forma, as regiões espectrais podem ser definidas em função do comprimento de onda que abrangem, conforme mostrado na Figura 2.1.

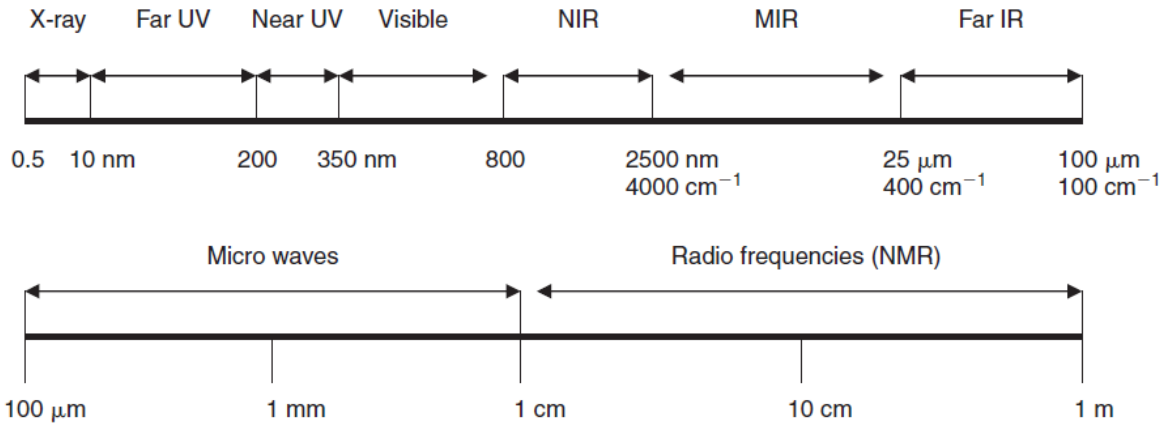

Figura 2.1: Regiões do espectro eletromagnético. Fonte: Sun (2009).

A radiação eletromagnética pode ser convenientemente representada como uma associação entre um campo magnético e um elétrico, que sofrem oscilações senodais em fase e na direção de propagação. O campo elétrico é o principal responsável pelos fenômenos relacionados à análise instrumental, tais como transmissão, reflexão, refração e absorção. O componente magnético, no entanto, é importante para a ressonância magnética nuclear, pois é responsável pela absorção de ondas de radiofrequência (Skoog, Holler e Crouch, 2007).

Enquanto a região do raio-x (comprimentos de onda entre 0,5 e 10 nm) está relacionada a variações de energia de elétrons das camadas internas de átomos e moléculas, a região do ultravioleta distante (10 - 200 nm) corresponde à emissão eletrônica a partir de orbitais de valência. Na faixa do ultravioleta próximo (200 - 350 nm) são observadas transições eletrônicas nos níveis energéticos desses orbitais. Nesse intervalo de energia, pode-se observar também a ocorrência da luminescência, ou seja, fluorescência e fosforecência (Sun, 2009).

A região da luz visível (350-800 nm) é outra fração do espectro onde ocorrem transições eletrônicas. Moléculas com um grande número de ligações duplas conjugadas absorvem energia nessa região (Sun, 2009).

O espectro infravermelho (IR) é usualmente subdividido em três regiões, denominadas IR-próximo, IR-médio e IR-distante, de acordo com os tipos de aplicações e de instrumentação. As técnicas e as aplicações de métodos baseados nessas três regiões diferem-se consideravelmente e, portanto, são normalmente tratadas separadamente (Skoog, Holler e Crouch, 2007). Uma vez que o IR-distante é particularmente útil para estudos inorgânicos, o que foge ao escopo deste trabalho, o mesmo não será abordado nos subcapítulos seguintes.

A região do infravermelho próximo (800–2500 nm ou 12500–4000 cm^{-1}), ou NIR, é a primeira fração espectral capaz de apresentar bandas de absorção relacionadas a vibrações moleculares. Caracteriza-se por bandas harmônicas e de combinação, que serão abordadas posteriormente. Essa região é amplamente utilizada na análise de composição de produtos alimentares.

A faixa do infravermelho médio (2500–25 000 nm ou 4000–400 cm^{-1}), ou MIR, é a principal região da espectroscopia vibracional, pois permite a identificação de moléculas orgânicas e a caracterização estrutural/conformacional de moléculas como proteínas, polissacarídeos e lipídeos.

Finalmente, na região de microondas (100 µm – 1 cm) a absorção de energia deve-se à rotação molecular. A faixa de radiofrequência (1 cm–10 m) é objeto de investigação da ressonância nuclear magnética (NMR) e ressonância eletrônica de spin (Sun, 2009). A Tabela 2.1 resume as faixas de frequências relevantes para propósitos analíticos, indicando também o método espectroscópico associado a cada uma delas. A última coluna da tabela lista os tipos de transições quânticas nucleares, atômicas ou moleculares que servem de base para cada técnica.

Tabela 2.1: Métodos espectrométricos comuns baseados na radiação eletromagnética.
Fonte: Adaptado de Skoog *et al*. (2007)

Tipos de espectroscopia	Faixa de comprimento de onda usual*	Faixa de número de onda usual, cm^{-1}	Tipo de transição quântica
Emissão de raios gama	0.005 – 1.4 Å	-	Nuclear
Absorção, emissão, fluorescência e difração de raios X	0.1 – 100 Å	-	Elétrons internos
Absorção de ultravioleta de vácuo	10 – 180 nm	1×10^6 a 5×10^4	Elétrons ligados
Absorção, emissão e fluorescência no ultravioleta – visível	180 – 780 nm	5×10^4 a 1.3×10^4	Elétrons ligados
Absorção no infravermelho e espalhamento Raman	0.78 – 300 µm	1.3×10^4 a 3.3×10^1	Rotação/Vibração de moléculas
Absorção de microondas	0.75 – 375 mm	13 – 0.03	Rotação de moléculas
Ressonância de spin eletrônico	3 cm	0.33	Spin de elétrons em um campo magnético
Ressonância Magnética Nuclear	0.6 – 10 m	1.7×10^{-2} a 1×10^3	Spin de núcleos em um campo magnético

* 1 Å=10^{-10} m=10^{-8} cm 1 nm = 10^{-9} m = 10^{-7} cm 1µm = 10^{-6} m = 10^{-4} cm

Nos métodos espectroscópicos, o analito pode ser estimulado por calor, energia elétrica ou reação química. A espectroscopia de emissão envolve métodos nos quais o estímulo é o calor ou energia elétrica, enquanto a espectroscopia de quimiluminescência refere-se à excitação do analito por meio de reação química. Em ambos os casos, a medida da potência radiante emitida quando o analito retorna ao seu estado fundamental pode fornecer informações sobre sua identidade e concentração. Os resultados dessas medidas são geralmente expressos em um gráfico denominado espectro, que relaciona a radiação emitida com a frequência ou comprimento de onda usado na excitação.

Quando uma fonte externa de radiação incide sobre uma amostra, diversos processos podem ocorrer. Entre eles, a radiação pode ser refletida, espalhada ou absorvida. Nesse último caso, a energia absorvida promove parte das espécies do analito para um estado excitado, conforme visto na Figura 2.2. Enquanto na espectroscopia de absorção mede-se a quantidade de luz absorvida em função do comprimento de onda, na espectroscopia de fotoluminescência (Figura 2.3) mede-se a emissão de fótons após a absorção. As formas mais importantes de fotoluminescência para as aplicações analíticas são as espectroscopias de fluorescência e fosforescência.

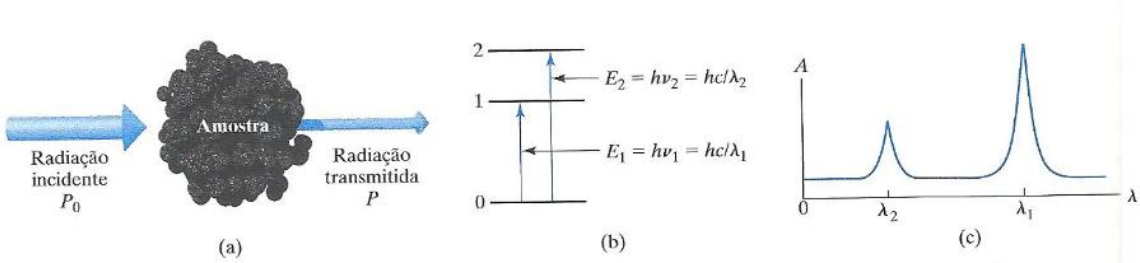

Figura 2.2: Métodos de absorção. A radiação incidente de potência *P0* pode ser absorvida pelo analito, resultando em um feixe transmitido com menor potência radiante P. Para que ocorra absorção, a energia do feixe incidente deve corresponder a uma das diferenças de energia mostradas em (b). O espectro de absorção resultante é mostrado em (c). Fonte: Skoog *et al*. (2007).

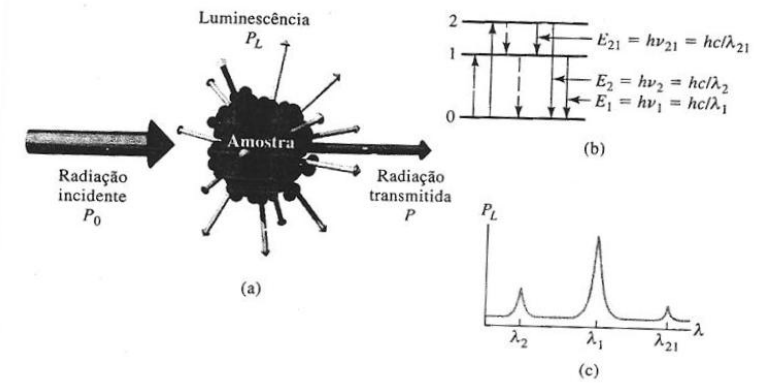

Figura 2.3: Métodos de fotoluminescência (fluorescência e fosforência). A fotoluminescência resulta da absorção de radiação eletromagnética seguida de dissipação de energia por emissão de radiação (a). Em (b), a absorção pode excitar o analito do estado 1 para o estado 2. Uma vez excitado, o excesso de energia pode ser perdido por

emissão de 1 fóton, causando a luminescência (representada pela linha contínua), ou por processos não-radioativos (linhas tracejadas). A emissão ocorre em todos os ângulos, e os comprimentos de onda emitidos (c) correspondem às diferenças de energia entre os níveis. Fonte: Skoog *et al*. (2007).

As técnicas de interesse para este trabalho são a espectroscopia de fluorescência bidimensional, o infravermelho (próximo ou médio) e a espectroscopia Raman. A seguir é apresentada uma breve revisão de cada um deles.

2.1.1 Espectroscopia de fluorescência bidimensional

O uso de fluorescência nas ciências biológicas sofreu um notável crescimento nas últimas décadas, se tornando atualmente uma técnica dominante, aplicada intensivamente em biotecnologia, citometria de fluxo, diagnósticos médicos, sequenciamento de DNA, análise forense, análises genéticas, análises ambientais, dentre muitas outras aplicações. Dessa forma, a espectroscopia de fluorescência e a fluorescência *time-resolved* são consideradas as principais ferramentas de investigação em bioquímica e biofísica. Dentre as vantagens, a detecção de fluorescência é altamente sensível e prática, uma vez que dispensa os custos e dificuldades de uso de rastreadores radioativos para a maioria das medições bioquímicas (Lakowicz, 2006).

Fluorescência é definida como um tipo de luminescência, ou seja, emissão de luz a partir de qualquer substância que ocorre devido a estados eletrônicos excitados. Além da fluorescência, pode-se falar também em mais duas categorias de luminescência: a fosforescência e a quimiluminescência, dependendo da natureza dos estados excitados que dão origem a cada fenômeno (Ferd, 1981).

Diferentemente da quimiluminescência, em que a excitação do analito se dá por reação química, tanto na fluorescência como na fosforescência a excitação é feita pela absorção de fótons. Como consequência, esses dois fenômenos são frequentemente denominados pelo termo mais geral fotoluminescência. A fluorescência difere da fosforescência pelo fato de as transições eletrônicas responsáveis pela fluorescência não envolverem mudança do spin eletrônico, ocasionando tempos de vida menores dos estados excitados. Usualmente, a fotoluminescência ocorre em comprimentos de onda maiores que os da radiação de excitação (Skoog, Holler e Crouch, 2007).

O fenômeno da fluorescência, na qual a espectroscopia de fluorescência é baseada, ocorre em estados excitados do tipo singletos, em que o elétron presente no orbital excitado é pareado com o segundo elétron (spin oposto) no estado orbital fundamental. Consequentemente, ao retornar ao estado fundamental, assumindo o valor determinado de spin, ocorre a rápida emissão de energia, na forma de fóton. Este fenômeno é tipicamente apresentado por moléculas aromáticas. Cabe salientar que, devido ao baixo tempo de emissão da fluorescência, a medição da mesma requer equipamentos óticos e eletrônicos sofisticados, devido à grande complexidade desta tarefa (Wang *et al.*, 2011).

A Figura 2.4 apresenta um diagrama parcial de níveis de energia (Diagrama de Jablonski) para uma molécula fotoluminescente típica. A linha horizontal mais grossa, na parte inferior do diagrama, representa a energia do estado fundamental da molécula, que é normalmente um estado singleto, e é denominada S0. À temperatura ambiente, este estado representa as energias da maioria das moléculas em solução.

As linhas grossas na parte superior são os níveis de energia para os estados fundamentais vibracionais de três estados eletrônicos excitados. As duas linhas à esquerda representam o primeiro (S1) e o segundo (S2) estados eletrônicos singleto. A linha à direita (T1) representa a energia do primeiro estado eletrônico tripleto. Como normalmente ocorre, a energia do primeiro estado excitado tripleto é mais baixa que a energia do correspondente estado singleto.

Diversos níveis de energia vibracional estão associados a cada um dos quatro estados eletrônicos, conforme sugerido pelas linhas horizontais mais finas. Como exemplificado na Figura 2.4, as transições de absorção podem ocorrer do estado eletrônico fundamental singleto (S0) para vários níveis vibracionais dos estados eletrônicos excitados singleto (S1 e S2). Vale salientar que a excitação direta do estado fundamental singleto para o estado excitado tripleto não é mostrada, uma vez que esta transição envolve uma mudança na multiplicidade, há uma probabilidade muito pequena da sua ocorrência. Uma transição de baixa probabilidade desse tipo é chamada transição proibida (Skoog, Holler e Crouch, 2007).

O retorno ao estado fundamental usualmente ocorre para um estado vibracional fundamental superior ao S0, o que rapidamente atinge o equilíbrio térmico. Este retorno a um estado excitado superior no nível de energia fundamental é o que dá origem à estrutura vibracional do espectro de emissão. Uma consequência interessante da emissão para níveis vibracionais de energia superiores a do nível fundamental é que o espectro de emissão é tipicamente um espelhamento do espectro de absorção da transição S0 para S1. Esta similaridade ocorre porque a excitação eletrônica não afeta de forma significativa a geometria nuclear, desta forma, o espaçamento entre os níveis de energia vibracional dos estados excitados são similares aqueles dos estados fundamentais. Como resultado, o espectro de emissão é similar ao espectro de absorção (Skoog, Holler e Crouch, 2007).

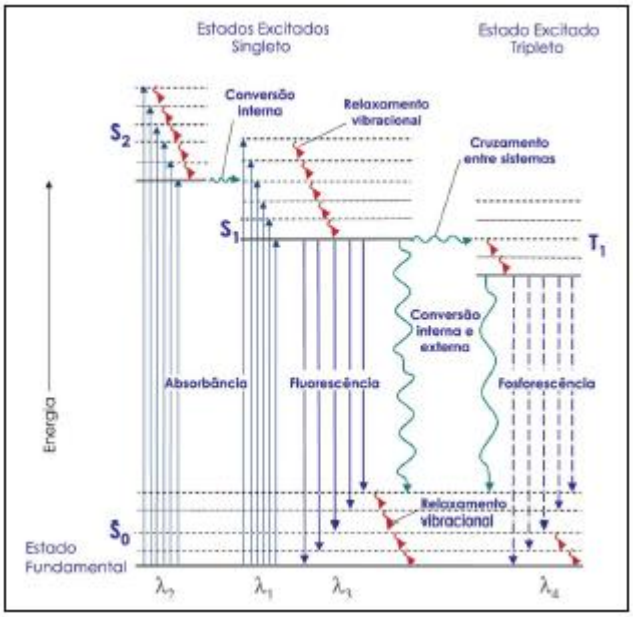

Figura 2.4: Diagrama parcial de níveis de energia para um sistema fotoluminescente.
Fonte: Sotomayor *et al.* (2008).

Dados espectrais de fluorescência são geralmente apresentados na forma de espectros de emissão. Um espectro de emissão de fluorescência é um gráfico da intensidade de radiação emitida em função do comprimento (nm) ou número de onda (cm^{-1}). A Figura 2.5 apresenta dois espectros típicos deste tipo, sendo o primeiro associado a uma amostra de perileno e o segundo à quinina. Espectros de emissão variam consideravelmente e dependem da química estrutural do fluoróforo e do solvente no qual ele está dissolvido (Lakowicz, 2006).

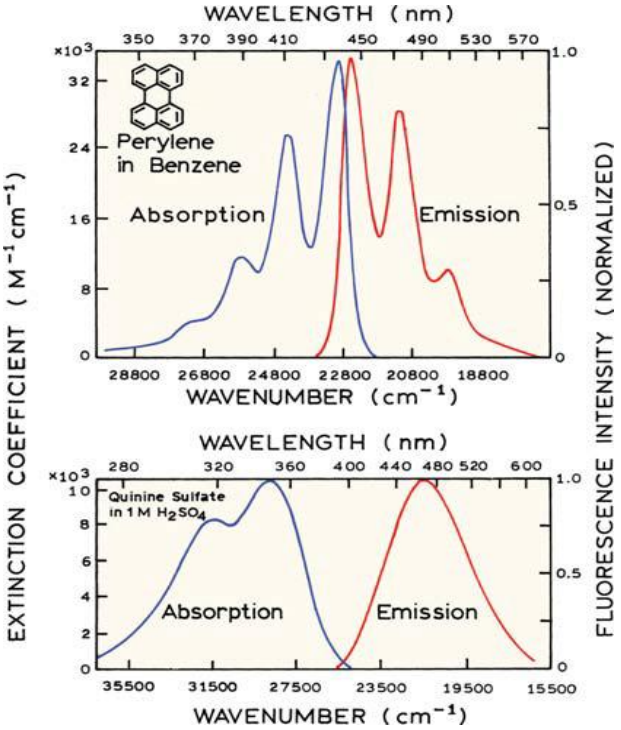

Figura 2.5: Espectros de absorção e emissão de fluorescência de perileno e quinina.
Fonte: Lakowicz (2006).

O espectro de absorção, ou excitação, é obtido pela medida da intensidade de luminescência em um comprimento de onda fixo, enquanto o comprimento de onda de excitação varia. Como a primeira etapa para gerar fluorescência é a absorção de energia radiante para a geração de estados excitados, um espectro de absorção é essencialmente idêntico a um espectro de excitação, obtido para as mesmas condições. Os espectros de fluorescência e fosforescência, por sua vez, envolvem a excitação em um comprimento de onda fixo enquanto é feito o registro da intensidade de emissão em função do comprimento de onda (Mulchandani e Bassi, 1995).

Na espectroscopia de fluorescência bidimensional, os dados espectrais são apresentados nos chamados espectro de luminescência total ou espectro fluorescente 2D, existentes na forma tridimensional ou em gráficos de contorno. Ambos gráficos mostram o sinal de luminescência em função dos comprimentos de onda de excitação e dos comprimentos de onda de emissão, de forma simultânea. O conjunto de dados que dá origem a este gráfico é usualmente chamado de Matriz de Excitação/Emissão. A Figura 2.6 apresenta um exemplo de cada um destes Espectros de Luminescência Total, onde em (a) é apresentado o espectro de fluorescência total de uma mistura de antraceno e de ovaleno, na forma de um gráfico em três dimensões, e em (b) é apresentado o gráfico de fluorescência total de 8-hidroxibenzopireno, na forma de curvas de contorno (Skoog, Holler e Crouch, 2007).

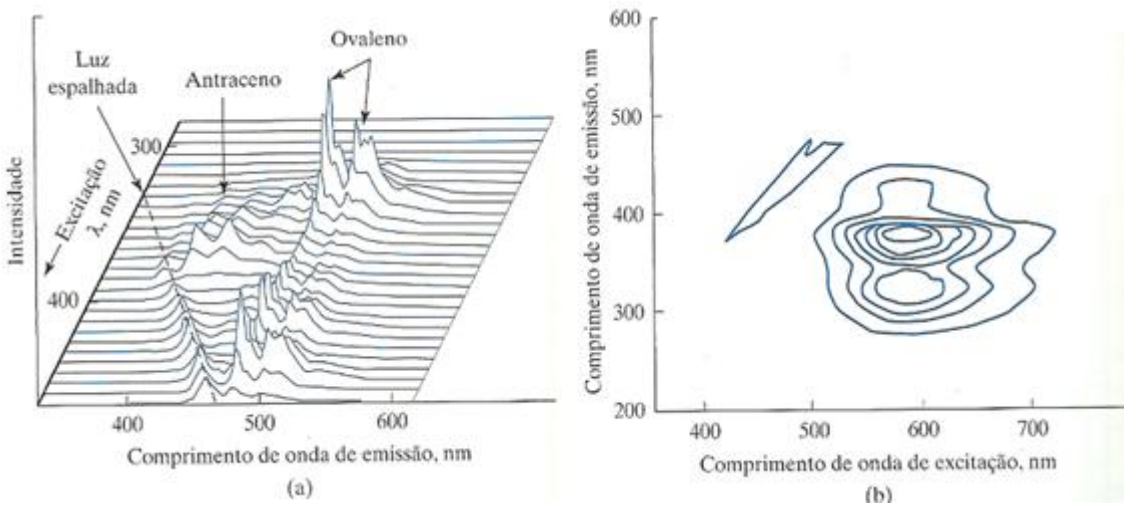

Figura 2.6: Espectros de luminescência total para (a) uma mistura de antraceno e ovaleno, e para (b) 8-hidroxibenzopireno). Fonte: Skoog *et al*. (Skoog, Holler e Crouch, 2007).

Alguns instrumentos de luminescência permitem varrer simultaneamente os comprimentos de onda de excitação e de emissão com uma pequena diferença de comprimentos de onda entre eles. O espectro resultante é conhecido como espectro síncrono. Um sinal de luminescência é obtido apenas em comprimentos de onda onde a excitação e a emissão ocorrem para a diferença de comprimentos de onda escolhida. O espectro síncrono também pode ser obtido através do espectro de luminescência total através de software apropriado.

Os componentes presentes em equipamentos destinados a medir fotoluminescência são similares àqueles encontrados em fotômetros ou espectrofotômetros ultravioleta-visível. A Figura 2.7 apresenta um esquema genérico do arranjo destes componentes nos fluorômetros ou espectrofluorômetros mais difundidos, ou seja, aqueles que empregam a ótica de duplo feixe em suas medidas. A principal diferença entre fluorômetros e espectrofluorômetros está no equipamento empregado para fazer a seleção dos comprimentos de onda de excitação e emissão: enquanto o primeiro usa apenas filtros, o segundo utiliza dois monocromadores para isolar os comprimentos de onda (Omary e Patterson, 1999).

Na técnica do duplo feixe mostrada na Figura 2.7, um dos feixes passa por um seletor de comprimento de onda de excitação e incide na amostra. A fluorescência emitida é isolada pelo seletor de comprimento de onda de emissão antes de atingir o transdutor. Enquanto isso, o segundo feixe, chamado de feixe de referência, passa por um atenuador antes de atingir o transdutor. Os componentes eletrônicos e o sistema computacional calculam a razão da intensidade de fluorescência para a intensidade do feixe de referência, que cancela o efeito das flutuações da intensidade da fonte (Skoog, Holler e Crouch, 2007).

O uso de espectroscopia de fluorescência 2D, aliado a métodos de análise de dados, é recorrente na literatura, com destaque para as aplicações online. Tal técnica mostrou-se

eficiente, por exemplo, no monitoramento online de cultivos, através da medição da reserva de NAD(P)H presente no interior de organismos. Através de dados de fluorescência, pode-se, portanto, inferir e controlar parâmetros de bioprocessos a fim de alcançar um alto rendimento de biomassa (Marose, Lindemann e Scheper, 1998). Além das aplicações em cultivos, esse tipo de espectroscopia tem sido aplicado na análise de reações quasi-enantioméricas catalisadas por enzimas proteases ou esterases (Knüttel *et al.*, 2001) e na análise de biotransformações dependentes de NADH catalisadas por oxidoreductase (Oliveira *et al.*, 2008).

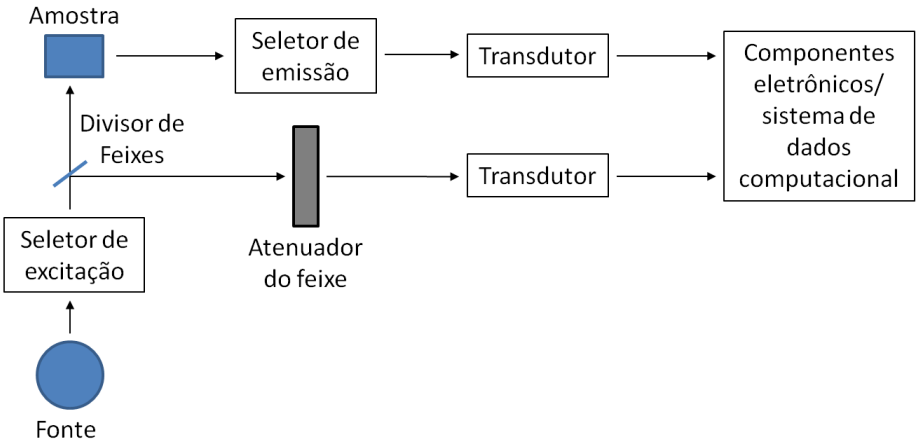

Figura 2.7: Esquema da estrutura genérica dos componentes de um equipamento fluorômetro ou espectrofluorômetro. Fonte: (Skoog, Holler e Crouch, 2007).

2.1.2 Espectroscopia no Infravermelho Médio (MIR) e Próximo (NIR)

A técnica de espectroscopia no infravermelho é classificada como uma técnica de espectroscopia vibracional, na qual também é incluída a técnica de espectroscopia Raman. Praticamente qualquer espécie que contenha ligações covalentes, orgânica ou inorgânica, absorve frequências de radiação eletromagnética na região do infravermelho. Da mesma forma que em outros tipos de absorção de energia, esse processo ocorre de forma quantizada e excita moléculas a um estado mais alto de energia. Do ponto de vista espectroscópico, as técnicas mais utilizadas para caracterização de compostos orgânicos utilizam as sub-regiões do infravermelho médio (MIR) ou próximo (NIR).

A Figura 2.8 exemplifica um espectro típico obtido com um espectrofotômetro comercial IR. A abscissa neste espectro é linear no número de onda (cm^{-1}), porém uma escala de comprimento de onda também é mostrada no topo do gráfico.

Cabe salientar que nem todas as ligações são capazes de absorver energia nessa região, uma vez que para isso é necessário a existência de um momento dipolo variante no tempo. Quando uma molécula polar vibra, ocorre uma oscilação regular em seu momento de dipolo, gerando um campo que pode interagir com o campo elétrico associado à radiação. De maneira similar, a rotação de moléculas assimétricas em torno de seus centros de massa resulta em oscilações periódicas do momento de dipolo, possibilitando interação com o campo de radiação. Assim, espécies homonucleares não podem absorver radiação IR, uma vez que não ocorre variação no momento de dipolo durante a vibração ou rotação de suas moléculas (Paiva, Lampman e Kriz, 2001).

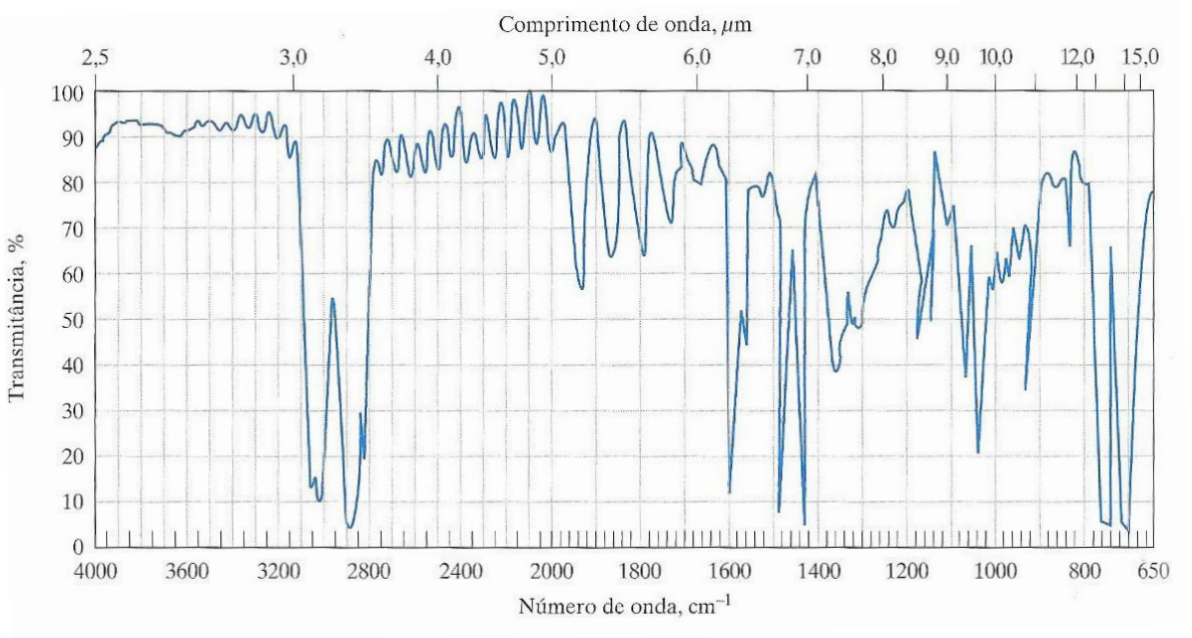

Figura 2.8: Espectro de absorção no IR de um filme fino de poliestireno. Fonte: (Skoog, Holler e Crouch, 2007).

Diversas são as vantagens de utilização da metodologia NIR e MIR para caracterização de processos. A Tabela 2.2 apresenta uma comparação entre as características qualitativas das técnicas de espectroscopia vibracional MIR e NIR. A partir desta tabela é possível escolher, baseado nas características almejadas no processo de caracterização, qual das técnicas apresenta melhor viabilidade para aplicação.

Tabela 2.2: Comparação das características qualitativas de MIR e NIR. Fonte: Adaptado de Pasquini (2002).

	MIR	NIR
	Vibrações Fundamentais	Sobretons e Combinações
Qualitativa	Excelente (estrutura)	Ruim (identidade)
Quantitativa	Excelente	Excelente
Intensidade	Alta	Baixa
Espessura da amostra	Muito pequena	Grande
Materiais	KBr/NaCl	Quartzo/Vidro
Sinal/Ruído	$<10^4$	$>>10^4$
Refletância	Satisfatória	Excelente

A região do MIR engloba as frequências de vibração de estiramento e de deformação angular das ligações da maioria das moléculas covalentes. Nesse processo, as frequências que correspondem àquelas de vibração natural da molécula em questão são absorvidas, e a energia associada causa um aumento da amplitude dos movimentos vibracionais das ligações moleculares (Skoog, Holler e Crouch, 2007). A Figura 2.9 ilustra os tipos de

vibrações moleculares do tipo estiramento e deformação angular capazes de absorver radiação IR.

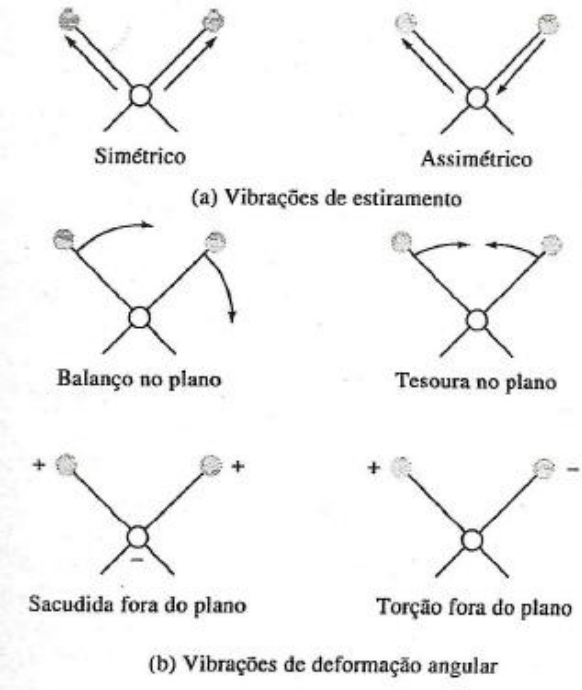

Figura 2.9: Tipos de vibrações moleculares de (a) estiramento e de (b) deformação angular. O sinal + indica movimento da página em direção ao leitor e o sinal – indica o oposto. Fonte: Skoog *et al.* (2007).

Na região do NIR, as bandas de absorção correspondem a sobretons ou combinações de vibrações fundamentais de estiramento que ocorrem na região de 3000 a 1700 cm^{-1}. As ligações envolvidas geralmente são do tipo C-H, N-H e O-H. Uma vez que as bandas são sobretons ou combinações, suas absortividades molares são baixas e os limites de detecção são da ordem de 0,1% (Skoog, Holler e Crouch, 2007).

A espectroscopia NIR é sensível de forma primária aos grupos funcionais das moléculas do analito e em menor escala a efeitos decorrentes dos níveis atômicos, microscópicos e macroscópicos das amostras. Com relação aos fatores químicos que afetam os espectros vibracionais, devem ser mencionados: a posição da banda de energia, que diz respeito a constante de força da ligação e massa dos átomos presentes nas moléculas, à intensidade da banda de energia, fazendo relação com a alteração do momento dipolo que acompanha as vibrações e por fim os fatores referentes à simetria das moléculas, uma vez que a simetria determina se a banda é classificada como ativa ou inativa, além de influenciar na probabilidade de se observar combinações e ressonância entre os modos vibracionais presentes. De forma genérica, a espectroscopia NIR é sensível a qualquer fator que afete a massa atômica, força de ligação, momento dipolo ou simetria da molécula (Burns e Ciurczak, 2001).

O método de espectroscopia no infravermelho próximo (NIRS – Near Infrared Spectroscopy ou simplesmente NIR – Near Infrared) possui algumas vantagens que motivam sua utilização na caracterização de processos, dentre as quais podem ser citadas a velocidade de medida quando comparada a outros processos (usualmente medidas são tomadas em menos de 1 segundo), sendo uma técnica não destrutiva e que na grande

maioria dos casos não necessita de preparação da amostra. Além destas vantagens, esta técnica é considerada altamente versátil. Caso a concentração do analito ultrapasse 1% da composição total, resultados promissores podem ser esperados quase que na totalidade dos casos (Burns e Ciurczak, 2001).

Estruturalmente, os equipamentos de espectroscopia NIR são similares aos instrumentos para medidas de espectroscopia UV-visível e MIR. Os componentes básicos são a fonte luminosa, detector e o elemento dispersivo, que pode ser um prisma ou então uma rede de difração. Equipamentos que trabalham com infravermelho por transformada de Fourier utilizam de interferômetros, especialmente para comprimentos de onda superiores a 1000 nm (Burns e Ciurczak, 2001).

Em equipamentos de espectroscopia NIR, usualmente, a fonte luminosa é uma lâmpada halógena de banda larga com janela de quartzo e aproveitamento de radiação entre 0,8 e 2,5 μm. A dispersão da radiação é feita com redes holográficas cortadas com laser e movidas por motores de movimento descontínuo (motor de passo). Diodos emissores de luz também estão sendo empregados para esta função. Além de apresentarem um tempo de vida útil elevada, possuem grande estabilidade espectral, além de reduzido consumo de energia (Burns e Ciurczak, 2001).

A espectroscopia de infravermelho, tanto MIR como NIR, tem sido aplicada extensivamente no ramo de combustíveis. Diversos trabalhos têm sido publicados neste aspecto, utilizando a espectroscopia para predição de diversas propriedades de diesel e/ou biodiesel (Baptista *et al.*, 2008; de Lira, de Albuquerque*, et al.*, 2010; de Lira, de Vasconcelos*, et al.*, 2010; Pilar Dorado *et al.*, 2011; Canha *et al.*, 2012; Zhang *et al.*, 2012), bem como para determinação do conteúdo de biodiesel em blendas combustíveis (Alves e Poppi, 2013).

2.1.3 Espectroscopia Raman

Na hipótese de espalhamento da radiação, sua interação com a amostra pode ser elástica (comprimento de onda da radiação espalhada é igual ao incidente) ou não-elástica. A espectroscopia Raman emprega o espalhamento não-elástico para produzir um espectro vibracional de moléculas. Nesse tipo de análise, a intensidade da radiação espalhada é registrada em função do deslocamento de frequência da radiação incidente (Skoog, Holler e Crouch, 2007).

Embora possam existir consideráveis similaridades entre os espectros Raman e os espectros de IR, existem suficientes diferenças entre os tipos de grupos que são ativos no IR e ativos no Raman para torná-las técnicas complementares, em vez de competitivas. Uma importante vantagem da espectroscopia Raman frente ao IR consiste no fato da água ser um solvente amplamente utilizado, uma vez que ela não altera o espectro obtido (Skoog, Holler e Crouch, 2007).

Os espectros Raman são adquiridos irradiando-se a amostra com uma potente fonte de laser de radiação monocromática visível ou tipo NIR. A radiação que incide na amostra produz espalhamento em todos os ângulos, O espectro da radiação espalhada é medido a algum ângulo, geralmente 90°, com um espectrômetro apropriado. A fim de evitar

fluorescência, os comprimentos de onda de excitação são geralmente afastados de uma banda de absorção do analito (Skoog, Holler e Crouch, 2007).

Conforme mostrado na Figura 2.10 (a), à medida que a radiação incidente de frequência ϑ_{ex} incide sobre a amostra, as moléculas da mesma são excitadas de um de seus estados vibracionais fundamentais para um estado de energia mais alta, chamado estado virtual, indicado pelos níveis com linhas tracejadas. Quando as moléculas relaxam, elas podem retornar ao estado vibracional inicial e emitir um fóton de energia $E = h(\vartheta_{ex} - \vartheta_v)$, onde ϑ_v é a frequência da transição vibracional. Alternativamente, se a molécula está no primeiro estado vibracional, ela pode absorver um quantum da radiação incidente, ser excitada a um estado virtual e relaxar de volta para o estado fundamental vibracional. Esse processo emite um fóton de energia $E = h(\vartheta_{ex} + \vartheta_v)$. Em ambos os casos, a radiação emitida difere em frequência da radiação incidente pela frequência vibracional da molécula ϑ_v (Skoog, Holler e Crouch, 2007).

A Figura 2.10 (b) apresenta um espectro Raman, tipicamente mostrado como uma função do deslocamento do número de onda $\Delta\bar{\vartheta}$, definido como a diferença, em números de onda (cm^{-1}) entre a radiação observada e a da fonte. É importante notar que a magnitude dos deslocamentos Raman é independente do comprimento de onda de excitação.

O espectro Raman consiste de duas regiões, referentes às emissões de frequências mais baixas, chamadas espalhamento Stokes, e às emissões de frequências mais altas, denominadas espalhamento anti-Stokes. Como geralmente o nível vibracional fundamental é mais densamente populado que os níveis excitados, as linhas Stokes são mais intensas do que as linhas anti-Stokes. A radiação espalhada elasticamente é de mesma frequência que o feixe incidente e é chamada de espalhamento Rayleigh (Skoog, Holler e Crouch, 2007).

No tocante às aplicações, assim como as espectroscopias de infravermelho, a espectroscopia Raman tem sido bastante utilizada na análise de diferentes tipos de combustíveis, incluindo gasolina (Li, J. e Dai, L., 2012; Li, S. e Dai, L.-k., 2012), querosene (Andrade *et al.*, 2003) e etanol (Mendes *et al.*, 2003).

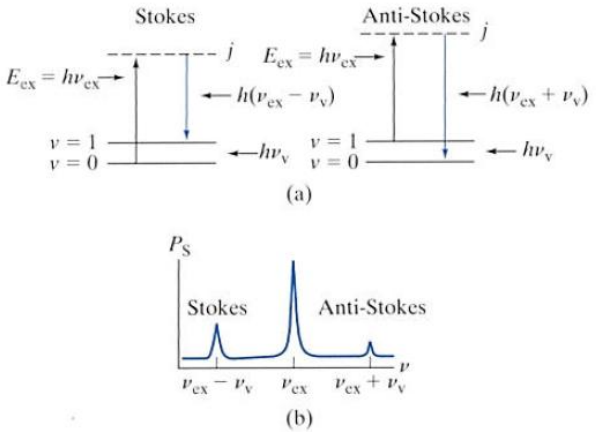

Figura 2.10: Origem dos espectros Raman. (a) Processo de reemissão de fóton de energia menor (esquerda) ou maior (direita). (b) Exemplo de um espectro Raman. Fonte: (Skoog, Holler e Crouch, 2007).

2.2 Métodos Quimiométricos

A característica comum entre as técnicas espectrométricas é que praticamente todas necessitam da análise de dados por métodos avançados para traduzir a grande quantidade de dados espectrais em informações úteis. Surge então a necessidade de uso da quimiometria, termo proposto há mais de 40 anos para descrever as técnicas e operações associadas à manipulação matemática e interpretação de dados químicos. Hoje, os métodos quimiométricos são amplamente reconhecidos como objeto de estudo e pesquisa para diversas aplicações envolvendo dados numéricos. Isso é ainda mais saliente nas ciências analíticas onde, devido à grande instrumentação, enfrenta-se uma quantidade enorme de dados que devem ser traduzidos em informação útil e de fácil compreensão. A quimiometria busca, então, aplicar ferramentas matemáticas e estatísticas a fim de auxiliar a aquisição, o processamento e a interpretação de dados analíticos (Adams, 1995).

A associação de técnicas instrumentais e métodos quimiométricos permite que os resultados analíticos sejam obtidos de forma sistemática e com confiabilidade estatística, permitindo reduzir o custo e o tempo dos experimentos. Através desta associação pode-se monitorar propriedades críticas durante determinado processo de fabricação, de modo a minimizar o risco sobre perdas de lotes (Adams, 1995).

2.2.1 Pré-processamento de dados

Na grande maioria dos casos, os dados de processo são desorganizados, ruidosos, incompletos, deficientes, altamente correlacionados, ou, muitas vezes uma combinação destes. Devido a estas não idealidades, a primeira etapa em qualquer análise de dados consiste no pré-processamento para ter acesso aos dados propriamente ditos e possivelmente melhorar a qualidade dos resultados (Beebe, Pell e Seasholtz, 1998).

Na prática, dados brutos provenientes de análises físico-químicas apresentam variabilidade proveniente de numerosas fontes (p.ex.: variações do meio reacional, ruído, interferências externas, etc). Nesses casos, os pré-tratamentos de dados mais aplicados são a (1) centralização dos dados na média, que consiste na subtração dos elementos de cada variável pela média desta no conjunto amostral, (2) escalonamento, onde cada elemento da variável é dividido pelo desvio padrão da mesma no conjunto amostral, de modo a equalizar a influência de cada variável no conjunto de dados e (3) auto-escalonamento, termo geralmente utilizado para indicar a centralização dos dados na média seguido de escalonamento (uso simultâneo dos procedimentos 1 e 2) (Wehrens e SpringerLink, 2011).

No entanto, tratando-se de dados espectrais, auto-escalonamento não é usualmente recomendado. Muito comumente, os dados consistem de áreas com alta quantidade de informação, contendo picos de diferentes intensidades e áreas contendo apenas ruídos. Quando cada variável espectral é normalizada ao mesmo desvio padrão, o ruído é convertido ao mesmo grau de importância dos sinais que contêm a informação atual. Esta situação é claramente indesejável e, nestes casos, a centralização na média é mais aconselhável (Ranzan, C., 2014).

Quando a intensidade espectral total é dependente da amostra, os espectros devem ser escalonados de forma que as intensidades possam ser comparadas. Além dos métodos de escalonamento acima apresentados, o método de normalização usualmente aplicado, especificamente em aplicações com NIR, é o SNV (Standard Normal Variate). Este método essencialmente faz o escalonamento nas amostras ao invés das variáveis. Isto é, cada espectro terá, após o escalonamento, média igual a zero e desvio padrão igual a um. Isto fornece dados livres de offsets e fatores de multiplicação. Obviamente, a suposição de que todo espectro deva ter a mesma média e variância não é sempre válida. Em certos casos, o fato de dados conterem intensidade superior em um espectro comparado aos demais pode conter informações importantes sobre o processo (Wehrens e SpringerLink, 2011).

2.2.2 Análise por Componentes Principais (PCA)

PCA é provavelmente a técnica quimiométrica não supervisionada mais difundida e, devido à importância de medidas multivariadas na química, ela é considerada por muitos como a técnica que mudou mais significativamente a visão de químicos sobre análise de dados (Brereton, 2003). Esta técnica tem por objetivo a redução da dimensão dos dados originais, facilitando a visualização das informações mais importantes.

Assim, constrói-se um novo sistema de eixos (denominados fatores, componentes principais, variáveis latentes ou ainda autovetores) para representar as amostras, sendo eles definidos de acordo com as direções de maior variabilidade dos dados. Dessa forma, a natureza multivariada dos dados pode ser visualizada em poucas dimensões.

O primeiro passo para a análise de componentes principais é a formação de uma matriz de variância/covariância (Z) dos dados (X) que irá isolar a fonte de variação dos dados. A matriz Z deve ser diagonalizável, devendo, portanto, ser simétrica, ou seja, Z = Z^T. Como qualquer matriz simétrica é ortogonalmente diagonalizável, então por uma transformação unitária tem-se:

$$\Delta = P^{-1}ZP = P^{T}ZP \tag{2.1}$$

em que Δ é uma matriz diagonal cujos elementos são os autovalores de Z e P é a matriz de autovetores, denominada *loadings* (pesos). Os autovetores devem ser ortogonais entre si. Basicamente, os *loadings* formam uma nova base ortonormal que explica os dados de X, sendo que a projeção dos dados nessa base é denominada *scores* (escores), representada pela matriz T. Desse modo, os dados são decompostos por um conjunto de vetores pesos e escores denominados componentes principais (PC) ou variáveis latentes (VL):

$$X = TP^{T} \tag{2.2}$$

Como a matriz $\mathrm{P^T}$ é ortonormal, $\mathrm{P^T P = I}$ (matriz identidade), portanto:

$$PX = TP^{T}P \tag{2.3}$$

$$T = XP \tag{2.4}$$

Em outras palavras, com uma mudança de coordenadas, o PCA tenta explicar a variação dos dados originais. O princípio básico do PCA está esquematizado na Figura 2.11.

Os eixos de componentes principais são os autovalores da matriz de covariância da matriz de dados observados e seus correspondentes autovalores indicam a proporção de variabilidade nas informações que cada autovalor leva em consideração. No PCA, a diagonalização da matriz de covariância resulta na decorrelação dos dados. Uma representação de dados de menor dimensão pode ser obtida selecionando um subconjunto de componentes principais com os maiores autovalores, para que em um dado número de dimensões, a representação através de PCA minimize o erro quadrático médio (Medeiros, 2009).

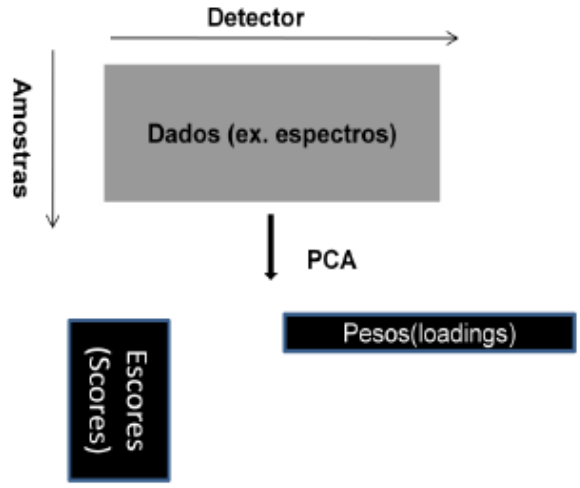

Figura 2.11: Princípios de PCA: decomposição dos dados espectrais em escores e pesos, formando o eixo de componentes principais que explica os dados originais. Fonte: Medeiros (2009).

Tendo definido as variáveis latentes, todas as amostras podem ser graficadas, ignorando os PCs de ordem maiores. Usualmente, poucos PCs são necessários para capturar a maior fração de variância do conjunto de dados (apesar disto ser altamente dependente da característica dos dados analisados)

A técnica de PCA possui inúmeras vantagens: é simples, possui uma única solução analítica e geralmente leva a uma representação dos dados mais simples de ser interpretada. A desvantagem desse método é que ele não produz, como resultado, um pequeno grupo de comprimentos de onda que carregam consigo a informação, mas sim um pequeno grupo de PCs, nos quais todos os comprimentos de onda estão representados (Geladi, 2003).

2.2.3 Modelagem Quimiométrica com Componentes Espectrais Puros (PSCM)

Dentre os métodos supervisionados aplicados a dados espectrais, uma alternativa à modelagem utilizando o espectro total é a modelagem quimiométrica com componentes espectrais puros, ou PSCM (Pure Spectral Chemometric Modeling), proposta por Ranzan *et al.* (2014). Essa metodologia é caracterizada pela utilização de modelos de regressão linear múltipla (MLR) associados ao método de otimização estocástica ACO (Ant Colony Optimization). Os elementos espectrais são filtrados, selecionando grupos de componentes espectrais com maior correlação com as variáveis de estado de interesse.

Após seleção do grupo de elementos espectrais, estes são utilizados como variáveis de entrada em modelos multilineares, calibrados e então submetidos a testes na predição das variáveis calibradas.

Assim, partindo de um conjunto de dados composto por amostras cujas variáveis de interesse estejam quantificadas, é realizada a análise de seus referidos dados espectrais para seleção do melhor conjunto de elementos capaz de descrever a variação destas variáveis dentro do conjunto amostral. Essa seleção é feita através da metodologia ACO, onde o grupo de elementos é selecionado através da minimização da função objetivo que quantifica o somatório dos erros entre os dados medidos e preditos das amostras do grupo de calibração. Para mais detalhes sobre esse método, ver o trabalho de Ranzan *et al.* (2014).

A fim de comparar o método PSCM com técnicas quimiométricas bem estabelecidas, tais como PCR e PLS, Ranzan *et al.* (2014) utilizou esses três métodos para gerar modelos de predição de conteúdo protéico de diferentes marcas de farinha. Resultados mostraram que PSCM atingiu uma acuracidade mais alta que PCR e PLS. Modelos utilizando essa nova metodologia apresentaram resultados melhores para qualquer tamanho de modelo, confirmando que a seleção e combinação de componentes espectrais puros fornecem algumas vantagens quando comparadas a metodologias que concentram todo conjunto de dados espectrais em componentes principais ou vetores pesos. Entre essas vantagens, pode-se citar o fato de não ser necessário medições completas do espectro, o que viabiliza a criação de pequenos sensores on-line baseados em espectroscopia e dispensa a normalização dos dados espectrais.

2.2.4 Seleção de variáveis

A construção de modelos empíricos por regressão linear frequentemente enfrenta o problema da seleção dos preditores, i.e., subconjunto de variáveis, mais relevantes dentre todas as variáveis disponíveis. Há várias razões para que o uso de um subconjunto reduzido de variáveis seja preferível frente ao uso de todos os dados disponíveis (Brown, Tauler e Walczak, 2009):

1. Uma redução do número de variáveis pode ser útil para diminuir o custo e tempo envolvido nas medições, tanto para procedimentos de calibração como para o subsequente uso rotineiro do modelo. Em aplicações envolvendo análises espectroscópicas, isso significa a possibilidade de desenvolver instrumentos mais práticos e versáteis para serem usados em campo, selecionando-se um número reduzido de comprimentos de onda dentro do espectro total.

2. Modelos com um número menor de variáveis preditoras tendem a ter uma interpretação física mais simples. Em espectroscopia, um modelo com poucos comprimentos de onda é mais facilmente compreendido em termos das características de suas bandas espectrais.

3. Se o modelo é construído utilizando-se Regressão Linear Múltipla (MLR), a incerteza das predições tende a aumentar com a relação K/N, onde K e N são o número de variáveis preditoras e observações utilizadas na regressão, respectivamente. Assim, o uso de muitos preditores pode gerar modelos com uma capacidade preditiva ruim. Esse problema pode ser ainda agravado pela presença de multicolinearidade entre os preditores, que deteriora o condicionamento da regressão.

Métodos de regressão baseados em variáveis latentes, basicamente *Principal Component Regression (PCR)* e *Partial Least Squares (PLS),* podem ser empregados para circundar o terceiro problema. No entanto, tais métodos não estão diretamente preocupados com os dois primeiros aspectos. Além disso, diversos autores têm apresentado evidências apoiando o uso de métodos de seleção de variáveis para melhorar a capacidade preditiva de modelos PCR/PLS (Spiegelman *et al.*, 1998; Leardi, Seasholtz e Pell, 2002; Goicoechea e Olivieri, 2003).

Não há uma única abordagem correta no que diz respeito à seleção de variáveis. A melhor metodologia irá depender de muitos fatores, tais como a dimensão do problema, a natureza estatística e física dos dados e os recursos computacionais disponíveis. Genericamente, uma estratégia de seleção de variáveis pode ser tratada como uma combinação entre critérios adequados para avaliar subconjuntos de variáveis e um algoritmo de busca que otimize tais critérios, tais como o algoritmo genético ou o algoritmo colônia de formigas.

A avaliação do subconjunto pode ser alcançada com ou sem a construção explícita do modelo de regressão. Por exemplo, as variáveis podem ser escolhidas com base na razão sinal/ruído individual, ou outra informação *a priori* disponível para o analista. Nesse caso, as equações de regressão não são invocadas. Alternativamente, dado um subgrupo de variáveis, pode-se conduzir a regressão e se utilizar procedimentos para analisar a qualidade do modelo resultante. Essas duas alternativas são conhecidas na literatura de aprendizagem de máquina sob os nomes de abordagens “filtro” e “wrapper”, respectivamente (Brown, Tauler e Walczak, 2009).

A prática mais comum na seleção de variáveis baseia-se nos próprios coeficientes de ajustes obtidos em regressões usando os dados, onde elementos que apresentam valores mais elevados são mantidos e aqueles cujo valor pode ser considerado estatisticamente insignificante são desprezados. Na etapa seguinte, os elementos previamente desprezados são acrescentados ao modelo e um novo ajuste é realizado, mantendo os elementos com coeficientes mais elevados. Este procedimento é repetido até que o número de variáveis desejado seja alcançado. Infelizmente, coeficientes de elementos altamente correlacionados assumem valores elevados, fazendo com que a ordem de grandeza dos coeficientes não seja um bom indicador de importância (Ranzan, C., 2014).

Usualmente, considera-se que o único modo efetivo de garantir a capacidade de modelagem de um pequeno grupo de variáveis é através do seu teste efetivo. No caso do modelo não apresentar resultados satisfatórios, a solução seria testar um novo grupo de variáveis. Assumindo alguma forma de quantificar o erro de medição, muitos pesquisadores vêm aplicando algoritmos de otimização na procura dos grupos de variáveis que fornecem a capacidade preditiva máxima para os referidos sistemas (Leardi, Seasholtz e Pell, 2002; Allegrini e Olivieri, 2011; Hemmateenejad *et al.*, 2011; Kumar *et al.*, 2014).

2.2.5 Métodos de Otimização Meta-heurística: Ant Colony Optimization (ACO)

Dentre os métodos de otimização meta-heurísticos conhecidos, os mais amplamente utilizados na seleção de variáveis são os do tipo Algoritmos Genéticos. A ideia principal

por trás destes métodos é tratar uma população de possíveis soluções como vetores (cromossomos) formados por valores binários (genes), e através da simulação do que seria a reprodução sexual, mesclar os melhores indivíduos. Desta forma, os indivíduos resultantes são formados por combinações das soluções contidas em seus pais. A qualidade dos descendentes é medida em uma etapa de avaliação, chamada de Aptidão (Fitness). Vetores com baixas aptidões não terão, ou então terão baixa probabilidade de reprodução, de forma que gerações subsequentes apresentarão cada vez melhores soluções. Este método obviamente imita o processo de seleção natural, onde apenas os indivíduos melhor adaptados têm a chance de se reproduzir, dando origem ao seu nome (Goldberg, 1989).

No entanto, a técnica de Otimização Colônia de Formigas (ACO – Ant Colony Optimization) tem sido recentemente aplicada para resolução de problemas envolvendo seleção de variáveis. ACO é um algoritmo de otimização concebido com base no comportamento coletivo de formigas quando saem em busca de fontes de alimento. Os indivíduos de um formigueiro executam tarefas coletivas e tomam decisões que requerem um alto grau de coordenação, tais como a construção do formigueiro, alimentação da ninhada, armazenamento e busca de comida e assim por diante (Allegrini e Olivieri, 2011).

A Figura 2.12 representa um experimento hipótetico onde as formigas “descobrem” de forma coletiva qual o menor caminho entre o formigueiro e a fonte de alimentos. Nele, uma fonte de alimentos é separada do formigueiro por uma trilha com diferentes caminhos possíveis. Inicialmente as formigas ocupam todos os possíveis caminhos entre o ninho e a fonte de comida, mas, ao final do experimento, todas as formigas ocupam somente o menor caminho entre os dois pontos. Considera-se que isto ocorre devido ao fato das formigas, quando em movimento, secretarem no solo certa quantidade de feromônio, marcando seu trajeto.

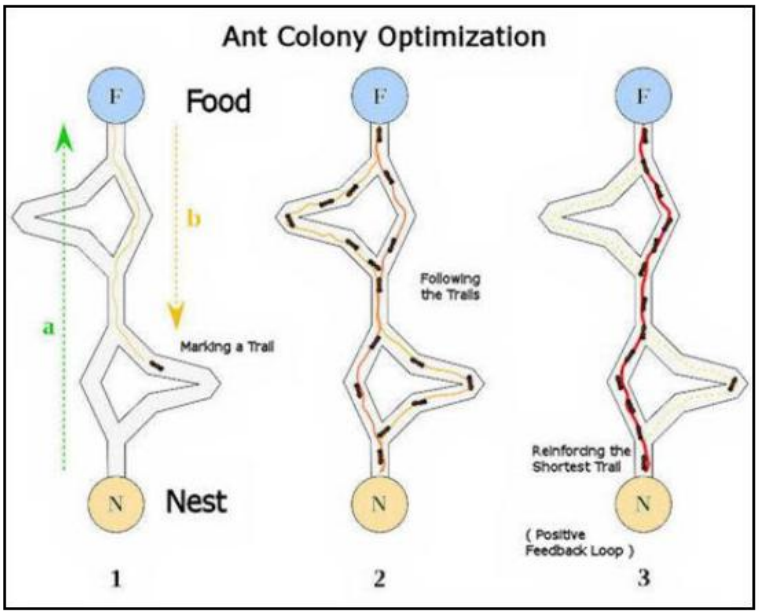

Figura 2.12: Diagrama mostrando a evolução no tempo do processo de busca de comida pelas formigas, onde o caminho entre o ninho e a fonte de comida é otimizado através do trabalho conjunto da colônia. Fonte: Adaptado de Goss *et al.* (1989).

Na teoria, o feromônio, por ser uma substância biologicamente ativa que evapora com o tempo, funcionaria como um chamariz para as demais formigas. As formigas que percorrem o menor caminho retornam ao formigueiro mais rapidamente, de forma que a trilha percorrida por estes indivíduos apresenta maior concentração de feromônio. Novas formigas que se encontrarem em bifurcações com esta trilha ótima, darão preferência à trilha com maior quantidade de feromônio (caminho mais curto e utilizado mais vezes), otimizando a distância percorrida entre o ninho e a fonte de comida.

Dorigo e Gambardela (1997) desenvolveram a primeira versão do ACO buscando solução para o problema do caixeiro viajante, um problema de pesquisa otimização combinatória no espaço de permutações. O algoritmo de Dorigo e Gambardella é fundamentado na distribuição de *m* formigas em *n* cidades de forma a permitir que cada uma das formigas percorra um caminho fechado passando uma única vez por cada uma das cidades e percorrendo o menor percurso para isso. As duas características fundamentais deste algoritmo são a forma como é feita a seleção das cidades e a simulação da vaporização e reforço da marcação de feromônio.

A variação da intensidade de feromônio é feita de forma que sua concentração em todas as trilhas ligando as cidades passe por dois processos de modificação:

- Diminuição da concentração, devido à multiplicação por um fator entre 0 e 1 (definido pelo usuário) após cada iteração, simulando a evaporação do feromônio no tempo.

- Aumento da concentração, sempre que a trilha é utilizada pelas formigas.

Partindo desses preceitos iniciais, diversas versões do algoritmo de otimização Colônia de Formigas foram desenvolvidas, dentre os quais se pode citar os trabalhos de Allegrini e Olivieri (2011), Hemmateenejad *et al.* (2011), Mullen *et al.* (2009), Socha e Dorigo (2008) e Ranzan, C. (2014).

A versão do ACO proposta por Ranzan, C. (2014) baseia-se nos mesmos preceitos do algoritmo original, mas é voltada para a busca de um grupo seleto de variáveis dentro do espaço disponível, de forma que apenas as mais indicadas são selecionadas, descartando as variáveis menos importantes. Além disso, ao contrário do problema original, a ordem de seleção não influencia o resultado. Essa nova versão utiliza o chamado “vetor de feromônios” para armazenar a marcação associada a cada variável disponível. Este vetor unidimensional, de tamanho igual ao número de elementos espectrais, é análogo à matriz de feromônios utilizada para seleção das cidades no algoritmo de Dorigo e Gambardella.

A Figura 2.13 apresenta um esquema do algoritmo implementado para resolução do problema de seleção de grupo de elementos espectrais utilizando ACO. O algoritmo está subdividido em quatro fases: inicialização das variáveis (fase 0), inicialização da solução (fase 1), rotina de busca (fase 2) e apresentação dos resultados (fase 3).

Na fase zero do algoritmo são definidas as variáveis necessárias para dar início à resolução do problema de otimização.

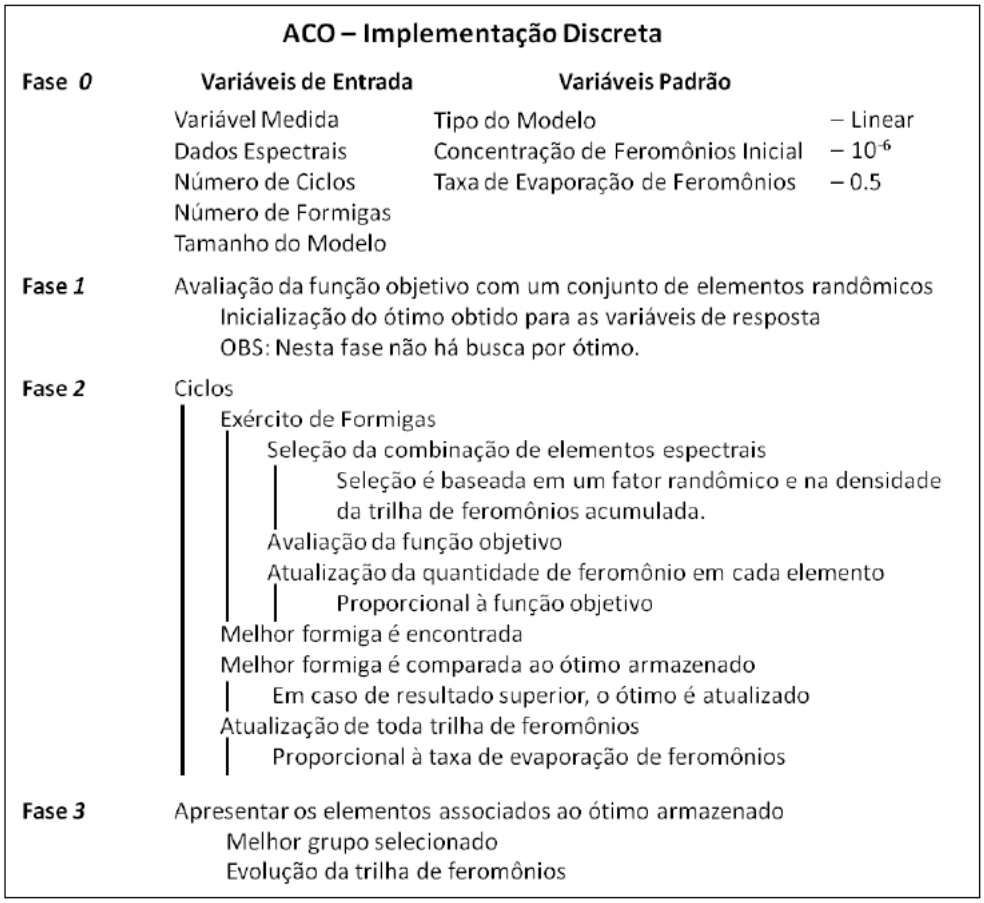

Figura 2.13: Representação esquemática do algoritmo ACO implementado por Ranzan, C. (2014) para seleção de grupos de elementos espectrais.

Na fase 1, é feita a inicialização do vetor de soluções, através da resolução da função objetivo com a seleção aleatória de um conjunto de elementos espectrais. Esta etapa tem o objetivo de unicamente iniciar o vetor de soluções de forma a se obter um objeto de comparação para os resultados obtidos da otimização na próxima etapa. Até esta etapa do algoritmo, nenhum tipo de otimização é realizado, as variáveis apenas são inicializadas.

O núcleo do algoritmo de otimização encontra-se na fase 2. É durante esta etapa que é feita a busca pela combinação de elementos espectrais que melhor predizem a variável de estado de interesse. Esta busca é realizada pela proposição de diferentes possíveis soluções (grupos de elementos), cujos resultados são comparados com o melhor resultado obtido até o momento, armazenado no vetor de soluções.

Nesta etapa, durante cada ciclo do algoritmo o exército de formigas varre o universo de possíveis soluções na busca pelo grupo de elementos espectrais que forneça o menor valor para a função objetivo. A função objetivo, neste caso, trata-se da soma dos quadrados de erros (SSE) de cada amostra, ou seja, o somatório da diferença elevada ao quadrado entre os valores observados da variável de estado e o predito pelo grupo de elementos selecionado. Cada um dos indivíduos do exército de formigas escolhe um grupo de elementos e submete-o ao teste da função objetivo. Caso o resultado seja inferior ao previamente armazenado, este é então substituído e o novo melhor resultado toma lugar no vetor de soluções.

A construção do grupo de elementos espectrais realizada por cada formiga é feita ordenadamente, de forma que os elementos são escolhidos um a um. Cada elemento previamente escolhido por uma formiga é retirado do conjunto de possíveis opções para aquela formiga, durante aquele ciclo, retornando ao vetor de possíveis escolhas para as próximas formigas, e para ela mesma no ciclo seguinte. A seleção de cada elemento espectral constituinte do grupo de predição a ser testado é baseada em dois fatores, sendo um aleatório e um baseado no vetor de marcadores (vetor da trilha de feromônios).

A finalidade do fator aleatório é garantir que o algoritmo de busca não fique retido em possíveis mínimos locais, fazendo com que toda a região de busca seja avaliada. Na prática, este fator é implementado através de uma função que gera valores randômicos entre 0 e 1. Cada vez que um novo elemento deve ser selecionado e adicionado a um grupo em formação, o algoritmo aciona o "gatilho randômico" e utiliza seu resultado como fator de decisão para selecionar o próximo elemento constituinte do grupo.

O valor fornecido pelo gatilho randômico é comparado com o valor de densidade de feromônio acumulada (equação 2.2), de forma que o elemento que apresentar densidade de feromônio acumulada igual ou imediatamente inferior ao valor fornecido pelo gatilho randômico é inserido no grupo de elementos da solução.

Na estratégia original de Dorigo e Gambardella, a função de densidade de feromônios acumulada é obtida através do cálculo da densidade de feromônio relativa (equação 2.1) para cada elemento espectral em função do total de marcador presente nos elementos, que é então avaliada de forma cumulativa desde o primeiro até o último elemento espectral disponível (equação 2.2).

Nas equações 2.1 e 2.2, ρ_{Fi} indica a densidade de feromônio relativa do elemento espectral i, F_i indica a quantidade de marcador associado ao elemento espectral i, N indica o número de elementos espectrais constantes no vetor de elementos espectrais e C_{Fi} indica a densidade de feromônio acumulada desde o primeiro até o *i-ésimo* elemento espectral.

$$\rho_{Fi} = \frac{F_i}{\sum_{i=1}^{N} F_i} \qquad 2.1$$

$$C_{Fi} = \sum_{j=1}^{i} \rho_{Fj} \qquad 2.2$$

A título de exemplo, Figura 2.14 apresenta um procedimento de seleção de elementos espectrais utilizando o gatilho randômico, uma trilha de feromônios simulada (F_i) composta por 1000 elementos espectrais e a sua respectiva curva de densidade de feromônios acumulada (C_{Fi}). No exemplo, o gatilho randômico é acionado em 0,43, indicando a seleção do respectivo elemento espectral 670.

O exemplo apresentado demonstra a dinâmica de qualificação dos elementos espectrais ao longo do processo de otimização utilizando-se a trilha de feromônios. A curva de CF é sensível a valores elevados de marcadores, de forma que elementos

espectrais com altas concentrações de feromônio apresentarão elevada taxa de variação da curva CF em função dos elementos espectrais, ao contrário de elementos com baixos valores de F, cuja taxa de variação da curva CF é próxima de zero.

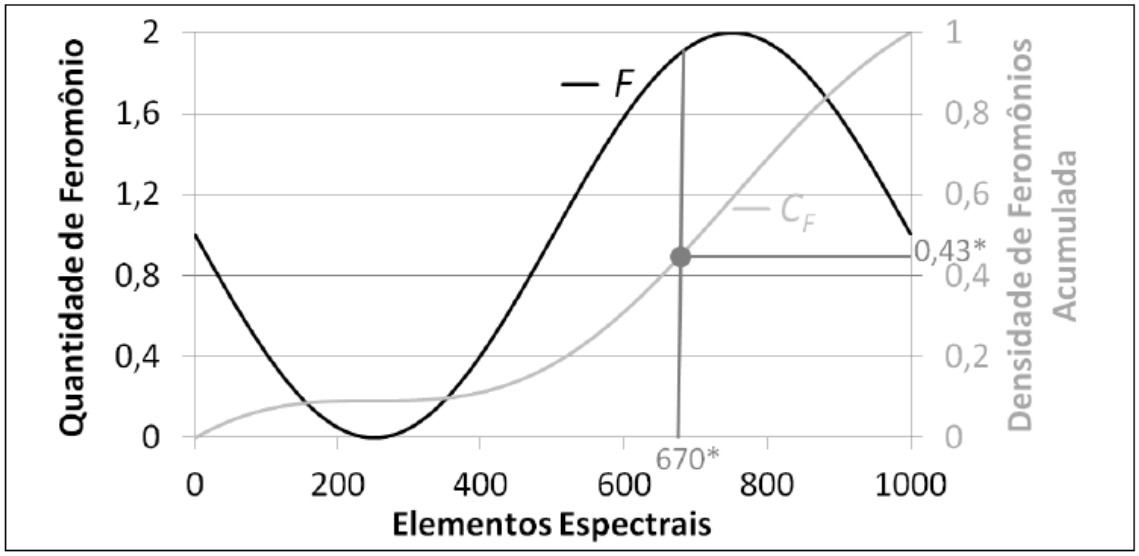

Figura 2.14: Exemplo de seleção de elementos espectrais utilizando a estratégia de Dorigo e Gambardela (1997).

O fato de o gatilho randômico ser utilizado para seleção de elementos associados aos valores da curva CF, e desta, por sua vez, apresentar variações significativas apenas para elementos com elevados valores de F, faz com que a probabilidade do algoritmo selecionar elementos com maior concentração de feromônios associadas seja elevada, priorizando assim a seleção dos elementos espectrais melhor avaliados pelo exército de formigas.

Após cada formiga selecionar e testar um determinado grupo de elementos, a quantidade de feromônios associada a cada elemento deste grupo é atualizada de forma inversamente proporcional ao somatório do erro quadrático entre os dados observados e preditos. Quanto maior o erro quadrático apresentado pelo modelo, menor é o incremento na quantidade de feromônios associada a cada elemento do grupo.

Esta estratégia faz com que elementos que apresentem maior correlação com a variável de interesse, ao serem adicionados a grupos de busca, produzam melhores resultados da função objetivo, aumentando o incremento da quantidade de feromônios nos elementos do grupo. Na evolução do processo de otimização, esta estratégia permite que os elementos espectrais sejam qualitativamente caracterizados em função de sua capacidade de predição da variável de interesse (Ranzan, C., 2014). No entanto, deve-se destacar o fato de que, como o critério SSE é uma propriedade do modelo como um todo, seu uso pode aumentar equivocadamente a concentração de feromônios de um componente espectral que, embora participe de um modelo com bons resultados, não constribui significativamente para os mesmos. Tal situação pode ser comparada ao que ocorre quando um grupo de alunos se reúne para realizar um trabalho escolar: mesmo que uns trabalhem mais que outros, todos recebem a mesma nota pelo resultado final.

Ao final de cada ciclo do procedimento de otimização, todo vetor de quantidade de feromônios é multiplicado pelo fator que simula a evaporação da trilha, penalizando

elementos que não foram selecionados pelo exército de indivíduos, no decorrer de cada ciclo.

Para finalizar o procedimento de seleção do grupo de elementos, na última etapa é apresentada a melhor solução obtida durante o processo de otimização. Este conjunto corresponde à melhor solução testada pelo exército de formigas no decorrer da rotina do algoritmo. Estes elementos, combinados através da estrutura de modelo proposta, são capazes de predizer os valores da variável de estado do processo apresentando o menor erro entre os dados preditos e medidos, sem a necessidade de manipulações nos dados de espectroscopia, ou mesmo coleta dos espectros completos (Ranzan, C., 2014). No entanto, sendo este um método estocástico, as soluções apresentadas variam a cada aplicação do algoritmo.

Ranzan *et al.* (2014) aplicaram este algoritmo ACO modificado para estimar o conteúdo de proteína em diferentes marcas de farinha com base nos dados espectrais de NIR. Os resultados mostraram que a utilização do ACO como ferramenta de filtragem tornou possível a seleção de regiões espectrais importantes, aumentando o coeficiente de determinação de modelos gerados em 60% quando comparados a outros métodos que utilizaram o espectro completo, tais como PCR.

Outros algoritmos de otimização também têm sido utilizados na seleção de variáveis, sendo que os dois algoritmos estocásticos mais conhecidos e aplicados no campo da quimiometria são o recozimento simulado e o algoritmo genético (Kirkpatrick, Gelatt e Vecchi, 1983; Cerny, 1985). Outros métodos, tais como a busca tabu, enxame de abelhas, enxame de partículas e busca harmônica podem também ser utilizados para esta aplicação (Mello e Pinto, 2008; Ghasemi-Varnamkhasti *et al.*, 2012).

2.3 Métricas de qualidade de modelos

Através das técnicas quimiométricas discutidas até aqui, é possível construir um grande número de modelos com base em dados experimentais, buscando-se predizer propriedades físico-químicas com precisão, exatidão e confiabilidade. Para isso, não basta apenas a obtenção de modelos quimiométricos: é necessário a utilização de critérios adequados para avaliar esses modelos, a fim de verificar sua qualidade de predição.

2.3.1 Raiz quadrada do erro médio de predição e calibração (RMSEP e RMSEC)

Dentre os parâmetros métricos úteis nessa avaliação, a raiz quadrada do erro médio de predição (RMSEP – Root-Mean-Square Error of Prediction) e o coeficiente de determinação R^2 são os mais utilizados. O RMSEP é calculado de acordo com a equação 2.3, onde $\hat{y}$ representa a variável predita pelo modelo, y a variável medida e n o número de amostras contidas no grupo de amostras de teste. Este critério analisa o ajuste do modelo ao conjunto de dados de teste avaliando a reprodutibilidade dos dados.

$$RMSEP = \sqrt{\frac{\sum_i^n (\hat{y}_i - y_i)^2}{n}} \qquad 2.3$$

Além destes, a raiz quadrada do erro médio de calibração (RMSEC) é a variação do RMSEP aplicada ao conjunto de dados do grupo de calibração, sendo crucial para a aplicação de rotinas de otimização que visam à diminuição deste como objetivo da função de otimização (Župerl *et al.*, 2011).

2.3.2 Soma dos quadrados dos erros (SSE) e coeficiente de determinação R^2

O coeficiente de determinação R^2 é uma medida da proporção da variabilidade dos dados originais explicada pelo modelo ajustado. Sua definição baseia-se na análise de variância, que utiliza os conceitos de soma dos quadrados dos erros (SSE, equação 2.4), soma dos quadrados de regressão (SSR, equação 2.5) e soma dos quadrados totais (SST, equação 2.6). Essas parcelas são dadas pelas equações a seguir, onde $\boldsymbol{e_i}$ é o erro amostral e $\overline{\boldsymbol{y}}$ é a média do vetor de dados medidos:

$$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad 2.4$$

$$SSR = \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2 \qquad 2.5$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad 2.6$$

O SSE representa a variabilidade não explicada pelo modelo, enquanto o SST representa a variabilidade total que deveria ser explicada pelo mesmo. Assim, a quantidade de variabilidade efetivamente prevista pelo modelo é dada por:

$$SSR = SST - SSE \qquad 2.7$$

O coeficiente de determinação R^2 pode, então, ser definido como:

$$\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \qquad 2.8$$

Esse coeficiente é usado com bastante frequência devido a sua simplicidade, porém apresenta algumas desvantagens em relação a sua interpretação. A confiança desse parâmetro, por exemplo, é uma função do tamanho do conjunto de dados de regressão. A adição de termos ao modelo diminui o SSE, consequentemente aumentando o R^2. Dessa forma, este é um critério perigoso para comparação de modelos concorrentes, uma vez que pode se tornar artificialmente alto devido a um eventual *overfitting* (pela inclusão de muitos termos no modelo).

2.3.3 Coeficiente de determinação ajustado (R_a^2)

O coeficiente de determinação ajustado (R_a^2) é uma variação do R^2 que pode ser utilizado para solucionar o problema abordado anteriormente, uma vez que leva em conta o número de graus de liberdade associado ao SSE e ao SST (Walpole *et al.*, 2012). Sendo k o número de termos do modelo:

$$R_a^2 = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{(n-1)}} \quad 2.9$$

2.3.4 Coeficiente de determinação adaptado (RR)

Além dos coeficientes de determinação tradicionais, pode-se citar ainda citar uma adaptação destes, denominada RR e dado pela equação 2.10 (Silveira, 2012), como uma forma de salientar diferenças na correlação R^2.

$$RR = -\log\frac{1}{1-R^2} \quad 2.10$$

2.3.5 Teste de hipótese t-student

Os conceitos utilizados na análise de variância são muito úteis em testes de hipóteses. O uso de testes t-Student de significância, por exemplo, é capaz de indicar a importância de cada parâmetro no modelo e é, portanto, muito útil na busca pelo modelo ótimo. Fazendo-se uso da estatística t é possível testar hipóteses a respeito dos coeficientes e construir seus respectivos intervalos de confiança com base na suposição de que as observações são amostradas aleatoriamente a partir de uma distribuição normal (Wilcox, 2012). Basicamente, as hipóteses a serem testadas são:

$$H_0: \beta_j = 0 \qquad H_1: \beta_j \neq 0 \quad 2.11$$

onde β_j é um determinado parâmetro j do modelo, sendo $\mathrm{j} = 0, 1, \ldots, \mathrm{k}$.

A rejeição ou não da hipótese H_0, denominada hipótese nula, depende do valor da estatística t. Esta, por sua vez, depende do nível de significância adotado, do parâmetro estimado e sua variância e do desvio padrão do erro.

Uma vez que a distribuição t-student é uma função do número de graus de liberdade, conforme visto na Figura 2.15, isso também deve ser levado em conta nos testes. Portanto, a hipótese H_0 não é rejeitada se o valor t obedece à inequação 2.12, onde $t_{\frac{\alpha}{2}}$ representa o ponto associado a uma probabilidade $\frac{\alpha}{2}$ em uma distribuição t-student com $n-k-1$ graus de liberdade.

$$-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}, \quad 2.12$$

Se isso ocorre, ou seja, se o coeficiente é considerado insignificante, conclui-se que tal variável explica uma quantidade insignificante de variabilidade da variável de saída na presença dos outros regressores do modelo.

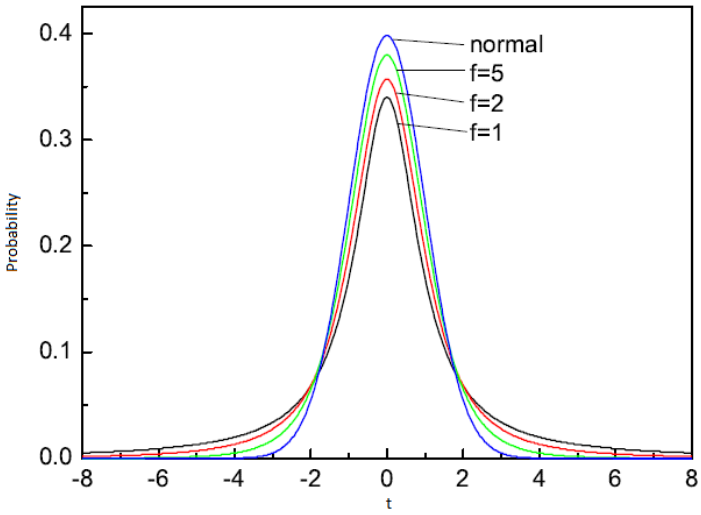

Figura 2.15: Distribuição t-student para 1, 2 e 5 graus de liberdade e distribuição normal. Fonte: Adaptado de Bohm e Zech [7070](70)[707070].

2.3.6 Intervalos de confiança

Mesmo considerando os estimadores de parâmetros como imparciais (valor esperado ou médio igual ao verdadeiro valor do parâmetro), é improvável que eles estimem os parâmetros β_j de forma exata. Embora a precisão da estimativa aumente com o tamanho da amostra, não é razoável esperar que uma estimativa pontual de uma determinada amostra seja exatamente igual ao parâmetro da população. Assim, é preferível determinar um intervalo o qual seja possível assumir, com determinada confiança, conter o verdadeiro valor do parâmetro β_j, chamado de intervalo de confiança.

Lembrando que a hipótese H_0 é válida e, portanto, o parâmetro é insignificante quando t recai no intervalo dado pela equação 2.12, é possível construir um intervalo de confiança para o parâmetro β_j, dado pela equação 2.13, capaz de estabelecer, com determinado grau de confiança, que o verdadeiro valor deste parâmetro encontra-se no seu interior. De forma geral, quanto menor o intervalo de confiança, ou seja, a diferença entre o limite positivo e o negativo, mais confiável é a estimativa daquele parâmetro (Walpole *et al.*, 2012).

$$\beta_j = b_j \pm t_{\frac{\alpha}{2}} \sqrt{\frac{SSE}{n-k-1}\, c_{jj}} \qquad 2.13$$

Na equação acima, b_j é o estimador do parâmetro e c_{jj} é a variância do parâmetro b_j, dada pela diagonal da matriz de covariância dos dados.

O tamanho do intervalo de confiança é, portanto, dado por:

$$LCI = 2 \times t_{\frac{\alpha}{2}} \sqrt{\frac{SSE}{n-k-1}\, c_{jj}} \qquad 2.14$$

2.3.7 Adaptação do teste de hipótese F

Outra possibilidade de avaliação da contribuição de cada preditor para o resultado é a utilizada de testes F para comparação de subconjuntos de variáveis contra o modelo completo. Isso é alcançado utilizando-se a equação 2.15, na qual SSE_{sub} e SSE são os erros obtidos por um submodelo e pelo modelo completo, respectivamente, e k_{sub} é o número de preditores que compõem o submodelo. Quanto maior o valor F, pior é o submodelo quando comparo ao original. Esta alternativa será melhor discutida no capítulo 3.

$$F = \frac{\frac{(SSE_{sub} - SSE)}{k - k_{sub}}}{\frac{SSE}{n - k - 1}} \qquad 2.15$$

A utilização de testes deste tipo tem a vantagem de avaliar a contribuição de cada parâmetro separadamente, ao invés da adequabilidade do modelo como um todo. Métodos de otimização podem, portanto, utilizar essas informações em associação com estatísticas do modelo para identificar e selecionar as variáveis mais relevantes.

2.4 Caracterização de Diesel

De um modo geral, um combustível é qualquer substância capaz de gerar energia que possa ser aproveitada para criar energia mecânica. Os tipos de combustíveis mais importantes no atual contexto mundial são os fósseis, chamados também de convencionais, uma vez que fornecem a maior parte da energia utilizada mundialmente, e os renováveis, devido ao crescente interesse em combustíveis alternativos capazes de gerar energia com menos impacto ambiental (Speight, 2008).

Produtos oriundos do petróleo são químicos altamente complexos, e é necessário um esforço considerável para caracterizar suas propriedades físico-químicas com um alto grau de precisão e acuracidade. A análise desses produtos é necessária para determinar as propriedades que podem ajudar na solução de problemas de processo, bem como as propriedades que indicam o funcionamento e desempenho dos produtos em serviço (Speight, 2002).

A determinação das propriedades de combustíveis, em especial diesel e biodiesel, utilizando-se métodos espectroscópicos tem fornecido resultados bastante promissores e apresenta muitas vantagens em relação à determinação por métodos tradicionais, tais como o fato de métodos espectroscópicos serem rápidos e não destrutivos. Vários são os trabalhos publicados sobre o assunto, a exemplo de Scherer *et al.* (2011) que, utilizando espectroscopia de fluorescência unidimensional (comprimento de onda de excitação é mantido fixo), estabeleceu com sucesso um método de determinação do conteúdo de biodiesel em blendas com diesel. Esse método baseou-se na diferença de absorção e de intensidade de fluorescência entre os dois componentes ao serem excitados com um comprimento de onda específico. O uso de um espectrofluorômetro unidimensional tornou a tarefa de determinação do percentual de biodiesel nas blendas mais prática e portátil, além de ser um método mais sensível que o tradicional baseado em espectroscopia no infravermelho.

Baptista *et al.* (2008) utilizaram espectros de NIR para determinação de modelos que estimam com erros aceitáveis o índice de iodo, o CFPP, a viscosidade cinemática a 40 ^{o}C e a densidade a 15 ^{o}C do biodiesel. Na região do infravermelho próximo, um componente geralmente absorve em mais do que um comprimento de onda e, em matrizes quimicamente complexas, a absorbância em um dado comprimento de onda pode ter contribuições de mais do que um analito. Assim, os dados foram avaliados usando análise de componentes principais (PCA) para uma análise qualitativa dos espectros, seguida pela regressão por mínimos quadrados (PLS) a fim de desenvolver os modelos de calibração entre os dados espectrais e analíticos. Vários métodos de pré-tratamento foram aplicadas aos espectros antes do desenvolvimento dos modelos e, com isso, foi possível uma predição das propriedades com erros similares aos métodos de referência.

Felizardo *et al.* (2007) publicaram outros dois artigos abordando o uso de NIR para predição de propriedades do biodiesel: um referente à determinação de ésteres metílicos e outro do conteúdo de metanol e água no biodiesel. A metodologia utilizada foi semelhante àquela publicada por Baptista *et al.* (2008), embora os modelos de predição de metanol e água tenham sido construídos com regressão por componentes principais (PCR) e PLS.

Balabin e Safieva (2011), por sua vez, obtiveram sucesso na classificação do biodiesel de acordo com sua matéria de origem (colza, girassol, coco, palma, soja, algodão, rícino, Jatrofa, linhaça ou óleo de cozinha usado) a partir de espectros NIR de cada amostra e quatro técnicas diferentes de análise de dados multivariáveis: regularized discriminant analysis (RDA), PLS, K-nearest neighbor (KNN) e suport vector mahcines (SVM). Obteve-se com isso um modelo classificador muito satisfatório utilizando-se KNN e SVM (erros de 6.2% e 5%, respectivamente), sendo inclusive indicado para uso industrial.

Em ordem de importância econômica, o diesel é o mais importante de todos os produtos petrolíferos (Parkash, 2010). Este faz parte da classe de produtos conhecida como “destilados intermediários”, ou seja, aqueles cujo ponto de ebulição é maior que o da gasolina e menor que o do gasóleo, abrangendo uma faixa de produtos com ponto de ebulição entre 175 e 375°C e número de átomos de carbono variando de 8 a 24.

As características do diesel dependem da natureza do petróleo, da origem, do processo de refino a partir do qual o combustível é obtido e do aditivo (se algum) utilizado (ASTM D-6258). Além disso, a especificação para o diesel pode existir numa variedade de combinações de propriedades, tais como volatilidade, qualidade de ignição, viscosidade, gravidade e estabilidade (Speight, 2002).

Nos primeiros anos de desenvolvimento dos motores a diesel, o aumento da potência e a confiabilidade eram os principais objetivos. Entretanto, nos últimos anos, o controle e a minimização das emissões se tornaram os principais objetivos no projeto de novos motores. As especificações de diesel estão sendo constantemente reformuladas para satisfazer os requisitos de baixas emissões e alta eficiência.

Segundo Parkash (2010), as especificações para o diesel são definidas com base, sobretudo:

1) No desempenho do motor, o qual depende, dentre outros, da viscosidade, das propriedades de fluxo a frio (ponto de névoa, CFPP e ponto de fluidez), número ou índice de cetano e índice Diesel.

2) No manuseio do combustível e durabilidade do sistema injetor, destacando-se como principais influências a lubricidade, as propriedades de fluxo a frio e o conteúdo de enxofre.

3) Na segurança e estabilidade do mesmo para depósito, dado, sobretudo, pelo ponto de flash ou fulgor;

4) No nível de emissões do motor afetando os padrões de qualidade do ar, sendo o teor de biodiesel e o conteúdo de enxofre as principais propriedades relacionadas a esse aspecto.

A próxima seção aborda os aspectos do conteúdo de enxofre no combustível diesel, uma propriedade bastante relevante para o quinto capítulo deste trabalho.

2.4.1 Conteúdo de Enxofre

O enxofre tem um efeito definitivo no desgaste de componentes de motores com ignição por compressão, tais como anéis de pistão. A presença desse elemento no combustível afeta diretamente a durabilidade do motor e dos seus componentes, além de contribuir para os depósitos na câmara de combustão e no sistema de injecção e interferir no funcionamento dos dispositivos de limpeza de gases de exaustão, tais como o filtro de partículas e conversor catalítico. Sob altas temperaturas de exaustão em motores pesados, a combustão de enxofre produz o ácido sulfúrico, o que aumenta a formação de material particulado. Por esta razão, a especificação de enxofre foi reduzida de 1000 ppm (em peso), no ano 2000, para valores entre 500 e 15 ppm em 2007 na maioria dos países desenvolvidos do mundo (Parkash, 2010).

No Brasil, o diesel vendido nas regiões metropolitanas e nas grandes cidades deve conter menos enxofre para diminuir a poluição urbana, enquanto nas áreas rurais e em pequenas cidades esse limite é maior. Em 2009, a especificação para o conteúdo de enxofre em diesel comercializado no interior era 1.800 ppm (S1800) e o metropolitano, no máximo, 500 ppm (S500). Em algumas regiões metropolitanas do Brasil, existia ainda a oferta do diesel com 50 ppm (S50) (CNT, 2012). A partir de 2013, no entanto, essa especificação baixou para 10 ppm em substituição ao S50 e 500 ppm em substituição o óleo diesel S1800. Desta forma, desde 2013, o Brasil usa óleo diesel rodoviário com especificação de enxofre igual a 10 e 500 ppm (ANP, 2013).

Diante disso, a aferição dos limites de enxofre presente no diesel comercial é uma certificação exigida em diversos países. Assim, foram desenvolvidos testes considerados padrões, que são aceitos mundialmente como confiáveis e precisos, para esta quantificação. Uma das instituições mais respeitadas na certificação destes testes é a Associação Americana para Testes e Materiais (American Society for Tests and Materials - ASTM), a qual rege grande parte dos testes exigidos pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) para produção e distribuição de óleo diesel no Brasil. Dentre as diversas normas técnicas publicadas pela ASTM acerca deste assunto, seis delas são classificadas como sendo mais importantes para a aferição da concentração de enxofre em amostras de diesel, sendo elas: ASTM D2622, ASTM D4294, ASTM D5453, ASTM D7039, ASTM D7212 e ASTM D7220.

As normas citadas anteriormente apresentam os testes certificadores de quantidade de enxofre em óleo diesel exigidos pela regulamentação da ANP estabelecida na portaria ANP 50/2013 para comercialização do mesmo no território brasileiro. A Tabela 2.3 apresenta um resumo do tipo de teste realizado e sua respectiva faixa de concentração detectável.

Além desses métodos padrões, existem também métodos ópticos de aferição com grande destaque no ramo industrial, sendo possível citar a espectroscopia de luminescência molar (incluindo medidas por fluorescência, fosforescência e quimiluminescência), a espectroscopia vibracional (espectrometria no infravermelho próximo, médio e distante), e as técnicas baseadas em espectroscopia Raman (Skoog, Holler e Crouch, 2007).

Na questão pela busca quantitativa de enxofre em combustíveis, a espectroscopia NIR se mostra muito efetiva para amostras com alta concentração de enxofre (mínimo 3500 ppm), obtendo resultados comparáveis aos métodos padrões de aferição regulamentados pela ASTM (Breitkreitz *et al.*, 2003). A espectroscopia MIR também tem mostrado excelentes resultados, como no trabalho de Ferrão *et al.* (2011), em que, através desta técnica, foi possível alcançar coeficientes de determinação altíssimos para os modelos de quantificação de enxofre em diesel, com concentrações entre 312 e 1351 ppm.

Entretanto, mesmo sendo possível utilizar espectroscopia NIR para classificar diesel que sofreu ou não forte hidrotratamento (Cramer *et al.*, 2009), as bandas de absorção emitidas por compostos organo-sulfurados são pouco significativas, e em baixas concentrações (<15 ppm) ainda não existem modelos ou métodos que utilizem esse tipo de espectroscopia capazes de predizer e quantificar satisfatoriamente enxofre total em combustíveis (Ranzan, L., 2014).

Tabela 2.3: Resumo dos testes ASTM apresentados e seus limites de detecção. Fonte: adaptado de Ranzan, L. (2014).

Metodo	Tipo de Teste	Faixa
D 2622	Espectroscopia fluorescente de comprimento de onda dispersivo de raios X	3 – 53 000 ppm
D 4294	Espectroscopia fluorescente de energia dispersiva de raios X	150 – 50 000 ppm
D 5453	Combustão da amostra; Fluorescência ultravioleta para quantificar dióxido de Enxofre	1.0 – 8000 ppm
D 7039	Espectroscopia fluorescente de comprimentos de onda monocromáticos dispersivos de raios X	3 – 2800 ppm
D 7212	Espectroscopia fluorescente de energia dispersiva de raios X usando contador proporcional de baixo ruído de fundo	7 – 50 ppm
D 7220	Espectroscopia fluorescente de energia monocromática dispersiva de raios X	3 – 940 ppm

O uso de espectroscopia de fluorescência na área de análise de combustíveis tem apresentado bons resultados, uma vez que diversos dos componentes de interesse emitem fluorescência, descartando a necessidade de preparação de amostra e possibilitando o desenvolvimento de pesquisas para construção de analisadores em linha (Riveros *et al.*, 2006; Pantoja *et al.*, 2011). Em trabalhos como o apresentado por Aburto *et al.* (2014) e por Ranzan, L. (2014), resultados promissores são discutidos para busca específica de enxofre em óleo diesel com o uso de espectros de fluorescência e métodos quimiométricos. No primeito trabalho, o autor demonstra que o uso de oxidação enzimática dos compostos sulfurados associado a medições de fluorescência permite a quantificação de conteúdos extremamente baixos (limite de quantificação igual a 3,7 ppb) de enxofre em diesel dessulfurizado. Utilizando uma análise de regressão simples, o autor conseguiu atingir um coeficiente R^2 igual a 0,95 entre o valor real de enxofre nas amostras e o valor predito pelo modelo. O trabalho de Ranzan, L. (2014), por outro lado, foi capaz de, utilizando o método PSCM, ajustar modelos baseados em pares de fluorescência capazes de predizer satisfatoriamente concentrações de enxofre em amostras de diesel S10, cuja concentração média é de 6,5 ppm.

Capítulo 3 – Algoritmo ACO modificado

Neste capítulo são apresentadas, de forma mais aprofundada, as variações do algoritmo ACO utilizadas neste trabalho, tendo como base os conceitos discutidos nas seções 2.2 e 2.3. Além disso, é apresentada a legenda utilizada para as diferentes versões (combinações de critérios C1 e C2) do algoritmo ACO implementadas, sendo esta de suma importância para a compreensão dos gráficos e das posteriores discussões de resultados.

O algoritmo implementado neste trabalho é uma modificação da versão proposta por Ranzan *et al.* (2014), apresentada na seção 2.2.5 e baseada na evolução da trilha de feromônios durante a análise de grupos de componentes espectrais. Este algoritmo aplicado à espectroscopia baseia-se na distribuição diferenciada de marcadores de ferômonio entre os vários componentes espectrais (preditores) disponíveis. Inicialmente, todos componentes são marcados com a mesma concentração de feromônios. O algoritmo ACO então seleciona componentes espectrais aleatórios para gerar um modelo, cuja qualidade é avaliada utilizando-se funções objetivo para a predição da variável de interesse. Com base nessa avaliação, o modelo é selecionado ou não e a concentração de feromônio associada a cada componente espectral incluído no modelo é atualizada. Para a seleção de novos grupos espectrais, o algoritmo associa o gatilho randômico com a densidade acumulada de feromônios de todos os preditores possíveis. Essa associação evidencia elementos significantes dentro do intervalo espectral analisado e, após um determinado número de iterações, um perfil de concentração de ferômonios se estabelece. Regiões com alta densidade de feromônios destacam aqueles componentes importantes na predição da variável de saída.

Conforme visto anteriormente na Figura 2.13, o algoritmo ACO utilizado como base possui 3 etapas principais, além de uma etapa inicial de definição das variáveis necessárias ao problema de otimização. O algoritmo proposto por este trabalho utiliza essa mesma estrutura de 4 etapas, porém sugere algumas modificações na implementação. A seguir, é descrito o funcionamento das 4 etapas que compõem o algoritmo utilizado neste trabalho. Para ver a implementação do algoritmo utilizando o software Matlab (Ver. 5.3.0.10183 R11, The Mathworks, Inc., Natick, USA), referir-se aos apêndices constantes neste trabalho.

3.1 Fase 0 – Inicialização das variáveis

Na fase zero do algoritmo são definidas as variáveis necessárias para dar início à resolução do problema de otimização. Nesta etapa são introduzidos treze parâmetros no programa, que são utilizados como entrada da função de otimização. A sintaxe dessa função de otimização, denominada "acow", em Matlab é a seguinte:

```
[combmin,erromin,parmin,iteracoes,trail] =
acow(Vdata,Sdata,nw,niter,nform,feromon,criterio,funcao,ro,model,Q,tau
0,selecao)
```

Assim, a função utiliza como entrada os seguintes dados ou parâmetros:

a) *Vdata:* O vetor coluna de variáveis observadas, onde cada linha corresponde à respectiva amostra dos dados de espectroscopia carregados;

b) *Sdata:* Os dados espectrais, dispostos na forma de matriz bidimensional, onde cada linha corresponde a uma amostra distinta e cada coluna corresponde a um determinado componente espectral;

c) *nw:* O tamanho do modelo, ou seja, que indica o número de elementos espectrais a ser buscado para construção do modelo (valor *default* igual a 3);

d) *niter:* A escolha do número de ciclos que o algoritmo realizará na busca pelo ótimo (valor *default* igual a 50);

e) *nform:* A escolha do tamanho do exército de formigas utilizado na busca, ou seja, o número de avaliações da função objetivo a cada ciclo (valor *default* igual a 100);

f) *feromon:* Valor que define se o algoritmo deve mostrar a evolução da trilha de feromônios ao final da otimização, sendo:

   0 - Evolução da trilha não é apresentada

   1 - Toda evolução da trilha é armazenada e apresentada

   2 - Apenas a última trilha é armazenada e apresentada

g) *criterio*: vetor contendo dois valores, o primeiro relacionado ao critério de atualização de trilha de feromônios e o segundo relacionado ao critério de seleção de modelos. Essa é a principal modificação sugerida por este trabalho e, portanto, será mais bem discutido na seção 3.5.

h) *funcao*: função utilizada para a construção dos grupos de variáveis utilizando o método definido em "selecao" e construção dos modelos com base no tipo de modelo definido em "*model*".

i) *ro*: A "taxa de evaporação de marcador", indicando o multiplicador do vetor de feromônios aplicado entre cada ciclo (*default* igual a 0,5, indicando que a cada ciclo todos os marcadores se reduzem à metade de seu valor anterior).

j) *model*: O tipo de modelo utilizado para avaliação do grupo de elementos espectrais, sendo por *default* do tipo linear sem interações, mas podendo assumir estrutura quadrática pura, quadrática com interações ou linear com interações;

k) *Q*: importância do erro (ou outro critério de qualidade do modelo encontrado) no ajuste (*default* igual a 100).

l) *tau0*: O valor inicial da trilha de feromônios, referente ao valor igualitário associado a cada elemento espectral antes do início do processo de otimização, de forma que todos os elementos iniciem com a mesma quantidade de marcador (valor default igual a 10^{-6});

m) *selecao*: valor que define a forma de escolha do próximo grupo de preditores a ser testado (*default* igual a 1), sendo:

1 – Roleta: utiliza o gatilho randômico para selecionar as variáveis que comporão o próximo modelo a ser testado.

0 – Torneio: seleciona as variáveis com a maior densidade de feromônios para compor o próximo modelo a ser testado.

Terminada esta fase inicial de definição das entradas, o algoritmo passa para a primeira fase do algoritmo propriamente dito: a inicialização do vetor de soluções.

3.2 Fase 1 – Inicialização do vetor de soluções

Nesta etapa, o vetor de soluções é inicializado utilizando-se uma avaliação qualquer da função definida pela entrada "*funcao*" na fase 0. Ou seja, é construído um grupo de variáveis aleatório e construído um modelo com este grupo. A função retorna então o valor do erro ou de outro critério de avaliação da qualidade do modelo encontrado, que será o ponto de partida do vetor de soluções. O próximo modelo, que será construído na Fase 2, terá seu desempenho comparado a este primeiro elemento e, com base nisto, este será substituído ou não.

Assim, como a função utiliza uma rota aleatória, ainda não é realizada qualquer tipo de otimização nesta etapa. A busca pelo ótimo ocorre na próxima etapa, denominada Fase 2.

3.3 Fase 2 – Otimização propriamente dita

Esta etapa é o núcleo de algoritmo, uma vez que nela ocorre a otimização propriamente dita. Nesta fase, primeiramente é feita a seleção dos grupos de variáveis utilizando o método "Roleta" ou "Torneio", conforme definido na Fase 0 do algoritmo. Em seguida, é construído o modelo formado pelas variáveis selecionadas e seu desempenho é avaliado.

A cada modelo construído ocorre a atualização da trilha de feromônios, em que a quantidade de feromônio associada a cada componente espectral compondo o modelo é imcrementada por um valor proporcional ao desempenho do mesmo. Além disso, a cada iteração é simulada a evaporação dos feromônios, de forma que suas quantidades diminuem devido à multiplicação por um fator igual a 0.5. Este procedimento é repetido conforme o valor estipulado na Fase 0 pelo parâmetro *nform.* Terminada esta etapa, a

próxima fase é a Fase 3, em que é feita a decisão de substituir ou não o modelo anterior pelo modelo atual com base no desempenho dos mesmos.

O desempenho dos modelos gerados pode ser avaliado de diferentes formas, que serão discutidas na seção 3.5, uma vez que esse ponto é a principal modificação sugerida por este trabalho.

3.4 Fase 3 – Comparação/seleção do modelo

Finalmente, a cada ciclo completo de iterações da Fase 2, o melhor modelo encontrado é comparado com o modelo armazenado. Se o desempenho do modelo encontrado é melhor que o armazenado, este é substituído e um novo ciclo de iterações da Fase 2 inicia. Esta sequência de etapas 2 e 3 ocorre de acordo com o número de ciclos definidos por *niter* na Fase 0.

O desempenho dos modelos gerados pode ser avaliado de diferentes formas, que serão discutidas na seção 3.5, uma vez que esse ponto é a principal modificação sugerida por este trabalho.

Finalmente, concluído todos os ciclos estipulados para o algoritmo, a função de otimização “acow” fornece 5 variáveis de saída:

a) *combmin:* componentes espectrais que formam o modelo com o melhor valor encontrado para a função objetivo (erro mínimo ou outro critério de avaliação da qualidade do modelo), também denominado modelo ótimo;

b) *erromin:* erro associado ao modelo ótimo;

c) *parmin*: parâmetros ajustados para o modelo ótimo;

d) *iterações: número de iterações necessárias para encontrar o modelo ótimo;*

e) *trail*: evolução da trilha de feromônios (apresentação depende do valor da entrada *feromon);*

3.5 Principais modificações do algoritmo

Analisando-se o funcionamento do algoritmo ACO, é possível destacar duas etapas principais em cada iteração, uma referente à atualização da trilha de feromônios associada a cada preditor (Fase 2) e outra referente à decisão de seleção ou não do modelo construído (Fase 3). Para ambas as etapas, é necessário estabelecer critérios adequados, não necessariamente iguais, que indiquem a qualidade do modelo em análise. Um resumo das etapas de implementação dentro do algoritmo ACO pode ser visto na Figura 3.1, em que o Critério 1 (C1) refere-se às métricas de atualização de trilha de feromônios e o Critério 2 (C2) refere-se às métricas de comparação/seleção de modelos.

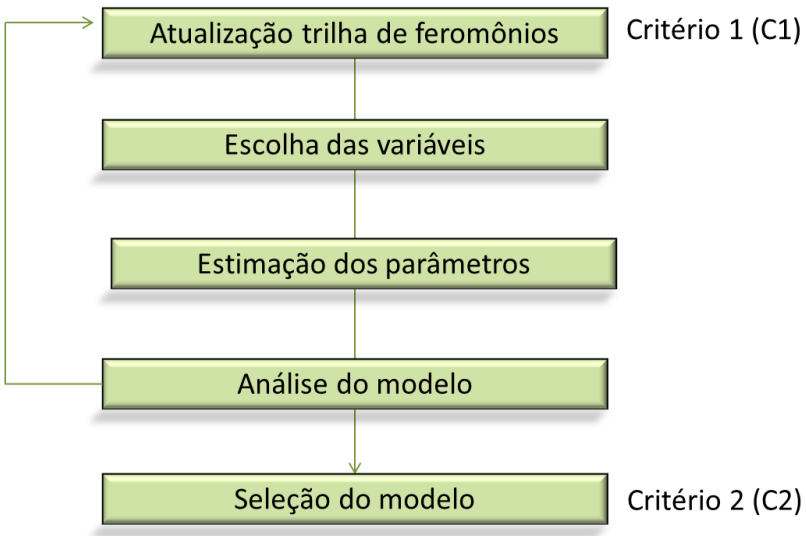

Figura 3.1: Resumo esquemático das etapas existentes no algoritmo ACO implementado neste trabalho.

Uma questão relevante é a escolha de um Critério 1 capaz de avaliar cada componente do modelo com base na sua contribuição para os resultados do mesmo. No algoritmo proposto por Ranzan, C. (2014), a distribuição da concentração de feromônios é feita com base na soma dos quadrados dos erros (SSE) do modelo. No entanto, tal critério não tem a habilidade de avaliar separadamente cada variável espectral que compõe o modelo, uma vez que apenas indica a qualidade de ajuste/predição do modelo como um todo. Assim, seu uso pode aumentar equivocadamente a concentração de feromônios de um componente espectral que, embora participe de um modelo com bons resultados, não constribui significativamente para os mesmos.

Dessa forma, além de um indicador da qualidade do modelo (global), é importante incluir no algoritmo métricas capazes de verificar a importância de cada preditor dentro do modelo (individual). A fim de estudar este aspecto, o presente trabalho propõe três métricas de avaliação do modelo e três métricas de avaliação dos preditores (associadas ao SSE) como critério para atualização da trilha de feromônios.

Além do SSE, as três métricas globais utilizadas são: o coeficiente de determinação ajustado (R_a^2), devido aos motivos discutidos na seção 2.3; o logaritmo desse coeficiente ($\log R_a^2$), a fim de enfatizar a região de interesse, ou seja, valores próximos a um; o valor absoluto de um coeficiente de determinação modificado, denominado RR e dado pela equação 2.10 (Silveira, 2012), como uma forma de salientar diferenças no coeficiente R^2.

Utilizando-se esses três critérios, o incremento de concentração de feromônio na trilha feito a cada iteração é exatamente o mesmo para todos os preditores compondo o modelo testado. Esse incremento é diretamente proporcional ao valor do R_a^2 e do RR e inversamente proporcional ao valor do SSE e do $\log R_a^2$.

As métricas individuais implementadas foram: o tamanho do intervalo de confiança do coeficiente associado a cada componente do modelo (o que pode ser utilizado como critério somente porque as variáveis são escalonadas); o teste t de hipóteses, atribuindo a cada componente espectral o valor de sua estatística t; o teste F de hipóteses, associando o valor F de um submodelo ao componente espectral ausente no mesmo. Estes três

critérios são divididos pela métrica global SSE, a fim de gratificar ou penalizar a contribuição de cada componente com base na qualidade do modelo.

Os dois primeiros critérios individuais seguem os conceitos expostos na seção 2.3. O teste F, por sua vez, funciona da seguinte maneira: primeiro, o algoritmo escolhe dentre todas as possibilidades um número k de variáveis, e um modelo de tamanho k é construído. Em seguida, um dos preditores é retirado do grupo escolhido originalmente e outro modelo, de tamanho $k_{sub} = k - 1$, é gerado. Esse submodelo é então comparado com o modelo completo através de um teste F. O valor de F obtido para esse submodelo é associado à variável **não** incluída no subgrupo. Portanto, variáveis importantes para o modelo final terão uma estatística F mais alta, uma vez que a sua não utilização gera modelos piores (alto F) que o modelo completo.

A cada iteração, as últimas três métricas apresentadas geram um aumento de feromônio diferente para cada componente espectral, levando em conta não só a qualidade do modelo, mas também a importância de cada preditor para o mesmo. Esse incremento é diretamente proporcional aos valores de t e de F, e inversamente proporcional ao tamanho do intervalo de confiança.

Para a comparação dos modelos gerados a cada iteração, métricas do tipo global devem ser utilizadas. Nesse aspecto, além do critério SSE proposto por Ranzan, C. (2014), foram considerados três métricas diferentes a fim de buscar o critério de seleção que mais enfatiza os melhores modelos. Os critérios escolhidos foram: a razão entre o SSE e o R_a^2; o produto entre o SSE e o $\log R_a^2$; a razão entre o erro e o coeficiente RR.

A Tabela 3.1 resume todos os critérios implementados e apresenta a legenda utilizada para cada combinação de critérios. Cada par de critérios (C1, C2), sendo C1 o critério de atualização de trilha de feromônios e C2 o critério de seleção de modelos, é representado por um número. Isso irá facilitar as discussões e a interpretação dos resultados ao longo deste trabalho.

Os diferentes critérios mostrados na Tabela 3.1 serão testados utilizando-se os estudos de casos discutidos no próximo capítulo. Cabe salientar que o caso número 1, que utiliza o SSE como C1 e C2, é aquele implementado por Ranzan, C. (2014), sendo utilizado algumas vezes como referência para os resultados produzidos neste trabalho.

A fim de fornecer uma referencia dos tempos computacionais de cada versão do algoritmo, a Tabela 3.2 apresenta o tempo necessário, em segundos, para cada uma das 28 versões resolver os problemas de otimização envolvendo os dois estudos de caso discutidos no próximo capítulo, utilizando um processador Intel® Core™ i5 750 @ 2.67 GHz com 8 GB de mémoria RAM. O primeiro problema consiste em encontrar o melhor modelo com 3 preditores dentre um universo de 150 componentes espectrais e 190 amostras, enquanto o segundo consiste em encontrar o melhor modelo com 3 preditores dentre um universo de 1150 componentes espectrais e 51 amostras. Assim, percebe-se que todas as versões do ACO necessitaram de tempos computacionais semelhantes, em torno de 340 segundos (5 minutos e 40 segundos) para resolver o problema de otimização.

Tabela 3.1: Resumo de todos os critérios implementados no algoritmo ACO para as etapas de atualização de trilha de feromônios (C1) e de comparação/seleção de modelos (C2).

		Seleção de modelos (C2)			
		SSE	SSE/R_a^2	$\|SSE * \log R_a^2\|$	SSE/RR
Atualização da trilha (C1)	SSE	1	2	3	4
	R_a^2	5	6	7	8
	$\|\log R_a^2\|$	9	10	11	12
	RR	13	14	15	16
	Tamanho intervalo de confiança (LCI)	17	18	19	20
	Teste t	21	22	23	24
	Teste F	25	26	27	28

Tabela 3.2: Tempo computacional necessário para resolução de dois problemas de otimização utilizando as 28 versões do algoritmo ACO.

	Tempo computacional (s)				
	Matriz espectral			Matriz espectral	
Versão	190 x 150	51 x 1150	Versão	190 x 150	51 x 1150
1	320	346	15	301	370
2	315	347	16	279	374
3	321	353	17	310	371
4	311	378	18	308	373
5	310	373	19	301	373
6	323	372	20	299	377
7	323	361	21	304	375
8	314	363	22	308	375
9	312	344	23	274	373
10	320	374	24	262	373
11	312	371	25	301	374
12	291	371	26	314	373
13	302	371	27	278	375
14	295	369	28	275	377

Capítulo 4 – Estudo de casos

Este capítulo apresenta os dois estudos de casos realizados com dados fornecidos por outros autores (Stärk *et al.*, 2002; Ranzan *et al.*, 2014), com o objetivo de avaliar as diferentes versões do algoritmo ACO sob dois aspectos: capacidade de otimização (encontrar o melhor valor dentre um conjunto de dados) e qualidade de inferência da variável de saída (prever um valor próximo ao real quando aplicado a um novo conjunto de dados). O primeiro estudo de caso tem como variável de saída a concentração de biomassa em um ambiente fermentativo de Saccharomyces Cerevisiar, em g⁄L, tendo como entrada dados espectrais provenientes da espectr oscopia de fluorescência bidimensional. O segundo estudo de caso tem como variável de saída o conteúdo de proteína em amostras de farinha, em g⁄kg, tendo como entrada dados espectrais provenientes da espectroscopia de infravermelho próximo.

A seção referente a cada estudo é composta por duas partes, uma que descreve o experimento e uma que discute os resultados. Como cada estudo utiliza diferentes materiais e métodos, optou-se por não criar uma única seção com este tema, mas introduzi-lo na subseção referente à descrição do experimento. Por último, são apresentadas as conclusões gerais do capítulo.

4.1 Fermentação Saccharomyces Cerevisiae

4.1.1 Descrição

Medidas baseadas na luz e em outras formas de radiação eletromagnética são amplamente utilizadas na caracterização de soluções e processos. Entre elas, a espectroscopia de fluorescência aparece como uma ótima técnica para o monitoramento online de cultivos, através da medição da reserva de NAD(P)H presente no interior de organismos. Essa espectroscopia baseia-se na emissão de fluorescência por fluoróforos presentes numa amostra devido à emissão/reemissão de luz de baixa energia. A luz reemitida possui sempre um comprimento de onda igual ou maior ao de excitação, e sua intensidade é proporcional à concentração de fluoróforos na amostra (Hitzmann *et al.*, 1998; Solle *et al.*, 2003).

Diante disso, o presente estudo de caso faz uso de dados de espectroscopia de fluorescência bidimensional de uma solução fermentada pela levedura *Saccharomyces cerevisiae* como variáveis de entrada para predição da concentração de biomassa da solução. Para isso, no entanto, é utilizado o algoritmo de otimização ACO como método de seleção dos componentes espectrais mais importantes, buscando-se encontrar o melhor modelo possível, considerando-se um número fixo de variáveis preditoras. Além disso, foram utilizadas diferentes implementações do algoritmo, aplicando-se os diversos pares de critérios (C1, C2) discutidos no capítulo anterior. Busca-se, com isso, estudar a diferença de desempenho de cada par de critérios na busca pelo melhor modelo.

Os dados experimentais de fluorescência utilizados neste trabalho consistem de duas bateladas de cultivo com fermentação de glicose da levedura *Saccharomyces Cerevisiae*, linhagem H620, cultivada em um biorreator de 1,5 L, a temperatura e pH constantes e iguais a 30 ^{o}C e 5,5, respectivamente, utilizando-se suplemento de meio Schatzmann. Uma batelada, denominada fermentação 1, é utilizada para fins de calibração e outra, denominada fermentação 2, para o teste do modelo.

Durante os cultivos, espectros de fluorescência foram coletados a cada 6 minutos pelo equipamento BioView-espectrômetro (Delta Light & Optics, Denmark), conforme descrito por Stärk *et al*. (2002). Cada espectro é constituído de 150 pares de comprimento de onda de excitação/emissão, compostos por 15 filtros na região de 270 nm a 550 nm (excitação) e 15 filtros na região de 310 nm a 590 nm (emissão), com largura de banda de 20 nm cada e coletados equidistantes por 20 nm. Os 150 pares de excitação/emissão são mostrados esquematicamente na Figura 4.1.

Uma vez que medições de fluorescência são feitas variando-se os comprimentos de onda de excitação e emissão, cada elemento da matriz de dados é composta por um comprimento de onda de excitação e outro de emissão. No entanto, o comprimento de onda emitido não pode ser menor que o absorvido, pois isso significaria uma energia de emissão maior que a absorvida, contrariando os princípios da física. Dessa forma, pares em que o comprimento de onda de emissão é menor que o de excitação não apresentam nenhuma informação real, apenas ruídos de medição.

Dessa forma, a matriz de dados de fluorescência, conforme apresentado na Figura 4.1, iguala a zero os valores de intensidade de fluorescência para aqueles pares localizados acima da diagonal principal. Valores válidos de intensidade, portanto, estão localizados abaixo da diagonal principal. Pares de fluorescência localizados na diagonal representam aqueles em o comprimento de onda de emissão é igual ao de excitação.

Foram coletados um total de 190 espectros de cada cultivo. Os dados obtidos pelo espectrofluorômetro BioView foram processados utilizando-se o software MATLAB (Ver. 5.3.0.10183 R11, The Mathworks, Inc., Natick, USA). Uma vez que a eficiência das metodologias de regressão está altamente associada com a qualidade dos dados espectrais, é importante normalizá-los previamente à análise. Tal processo ajuda a eliminar offsets e fatores multiplicadores e foi feito aplicando-se a normalização SNV (*Standard Normal Variate*) aos dados espectrais. A Figura 4.2 exemplifica um dos 190 espectros de fluorescência normalizados obtidos para a primeira e segunda fermentação.

Uma questão importante a ser avaliada nesses conjuntos de dados é a viabilidade de comparação entre os dois processos ao longo de todo o tempo de fermentação, a fim de separá-los em grupo de calibração e grupo de teste. A forma mais usual de avaliação qualitativa de dados de processo é através da análise de componentes principais. Aplicando PCA nos dois conjuntos de dados, previamente normalizados com SNV, e avaliando os Scores obtidos pode ser visualizado que os dados de fluorescência para os dois ensaios não apresentam diferenças que devam ser analisadas com maior aprofundamento. A Figura 4.3 apresenta o gráfico dos escores dos componentes principais (PC) 1 e 2, para os dois ensaios fermentativos.

Comprimento de onda de excitação (nm)																	
	Aberto	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	0
	550	0	0	0	0	0	0	0	0	0	0	0	0	0	0	134	135
	530	0	0	0	0	0	0	0	0	0	0	0	0	0	131	132	133
	510	0	0	0	0	0	0	0	0	0	0	0	0	127	128	129	130
	490	0	0	0	0	0	0	0	0	0	0	0	122	123	124	125	126
	470	0	0	0	0	0	0	0	0	0	0	116	117	118	119	120	121
	450	0	0	0	0	0	0	0	0	0	109	110	111	112	113	114	115
	430	0	0	0	0	0	0	0	0	101	102	103	104	105	106	107	108
	410	0	0	0	0	0	0	0	92	93	94	95	96	97	98	99	100
	390	0	0	0	0	0	0	82	83	84	85	86	87	88	89	90	91
	370	0	0	0	0	0	71	72	73	74	75	76	77	78	79	80	81
	350	0	0	0	0	59	60	61	62	63	64	65	66	67	68	69	70
	330	0	0	0	46	47	48	49	50	51	52	53	54	55	56	57	58
	310	0	0	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	290	0	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	270	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		310	**330**	**350**	**370**	**390**	**410**	**430**	**450**	**470**	**490**	**510**	**530**	**550**	**570**	**590**	**Aberto**
		Comprimento de onda de emissão (nm)															

Figura 4.1: Diagama mostrando os pares de fluorescência utilizados na aquisição dos dados espectrais, bem como o número associado a cada par. Fonte: Ranzan, C. (2014).

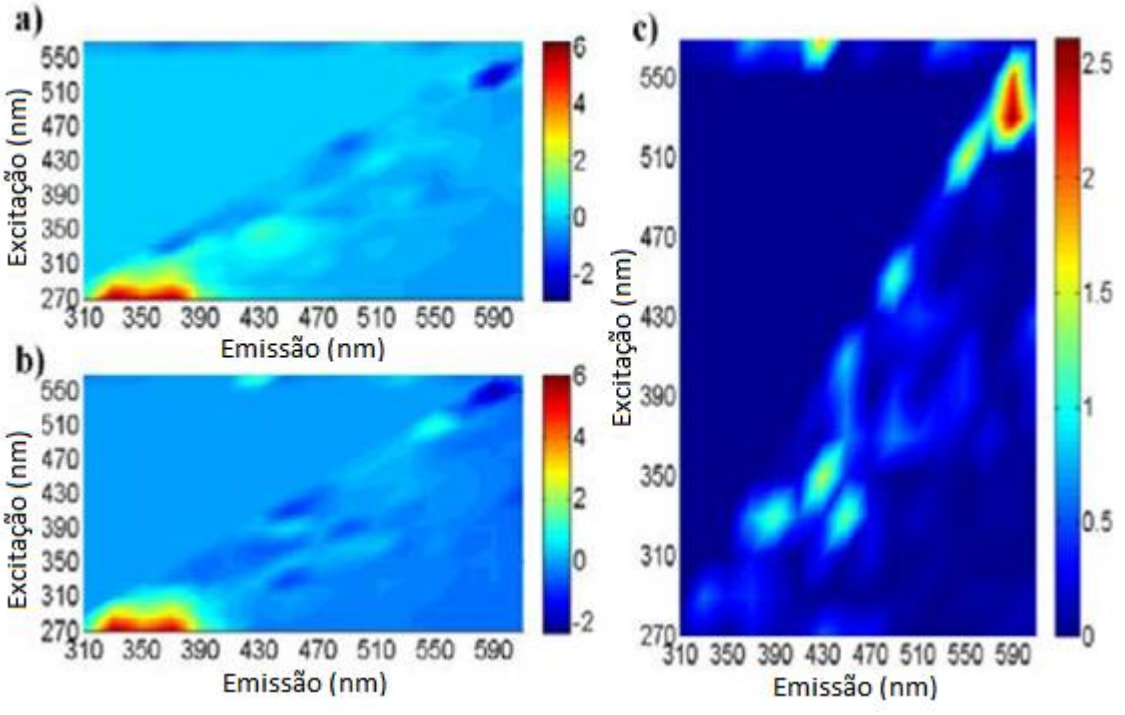

Figura 4.2: Espectro de fluorescência no tempo t=0, após aplicado o método SNV, da (a) fermentação 1 e (b) fermentação 2. (c) Diferença absoluta na intensidade de fluorescência, par a par, entre os espectros normalizados. Fonte: Ranzan, C. (2014).

Uma vez que os pontos correspondentes aos dois ensaios assumem valores próximos e evolução equivalente no decorrer dos ensaios, pode ser concluído que os ensaios foram conduzidos da mesma forma, tratando assim de dados espectrais comparáveis do mesmo processo bio-químico. Caso houvesse uma diferença significativa entre os resultados obtidos por PCA para os dois processos, seria necessário promover diferentes pré-tratamento dos dados espectrais, visando reduzir influências causadas por variáveis externas não controladas (Kara *et al.*, 2010).

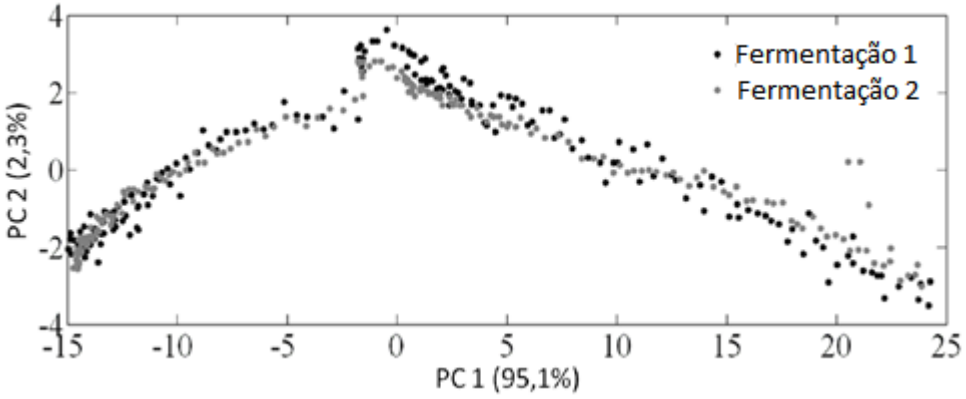

Figura 4.3: Componente principal 1 versus componente principal 2, para os dois ensaios fermentativos analisados. Fonte: Ranzan, C. (2014).

Como este não é o caso para o conjunto de dados em questão, é possível a sua utilização para fins quimiométricos apenas aplicando-se a normalização SNV. Porém, caso fosse necessário, os procedimentos mais usuais seriam normalizar os espectros, centrar os dados na média, derivar e suavizar utilizando o algoritmo de Savitzky-Golay e aplicar a correção de espalhamento multiplicativo (*MSC, Multiplicative Scatter Correction*).

Normalização é um tipo de pré-processamento que tem como objetivo reduzir a influência de variações indesejadas presentes no conjunto de dados, garantindo que cada observação seja representada de forma adequada e consistente.

Centrar os dados na média consiste em calcular a média das intensidades para cada comprimento de onda e subtrair cada uma das intensidades do valor médio. Desta maneira, cada variável passará a ter média zero, ou seja, as coordenadas são movidas para o centro dos dados, permitindo que diferenças nas intensidades relativas das variáveis sejam mais fáceis de perceber.

Deslocamento e inclinação de linha de base podem ser corrigidos por derivação dos espectros. Os métodos de alisamento são utilizados para reduzir matematicamente o ruído, aumentando com isto a relação sinal/ruído. Nestes métodos, é selecionada uma janela, a qual contém certo número de variáveis. Os pontos na janela são, então, utilizados para determinar o valor no ponto central da janela e, assim, o tamanho da janela influencia diretamente o resultado do alisamento. No método de Savitzky-Golay, um polinômio de ordem baixa é ajustado aos pontos da janela e utilizado para recalcular o ponto central.

A correção de espalhamento multiplicativo é um método de transformação utilizado para compensar os efeitos aditivos e/ou multiplicativos em dados espectrais. Este método remove a influência de efeitos físicos nos espectros, tais como o tamanho de partícula, a rugosidade e opacidade, os quais não trazem informações químicas sobre as amostras e introduz variações espectrais, como o deslocamento da linha de base. Para fazer a correção, o método MSC assume que cada espectro é determinado pelas características químicas da amostra somadas às características físicas indesejadas (Souza & Poppi, 2012).

Paralelamente à coleta de dados on-line por espectroscopia de fluorescência, os cultivos foram caracterizados de forma off-line por análise gravimétrica, de forma a possibilitar o acompanhamento da evolução das concentrações biomassa no meio fermentativo. No entanto, a fim de avaliar eficientemente modelos quimiométricos baseados em dados de fluorescência, a informação a respeito da variável de estado deve estar disponível com a mesma extensão amostral que os dados espectrais. Uma vez que, originalmente, a quantidade de dados off-line era consideravelmente menor que a de amostragens de dados espectrais, foi adotado um modelo dinâmico para o sistema fermentativo para interpolar a variável de estado, obtendo-se assim dados off-line e espectrais com a mesma amostragem. A simulação feita por esse modelo da evolução da fermentação 1 pode ser vista na Figura 4.4.

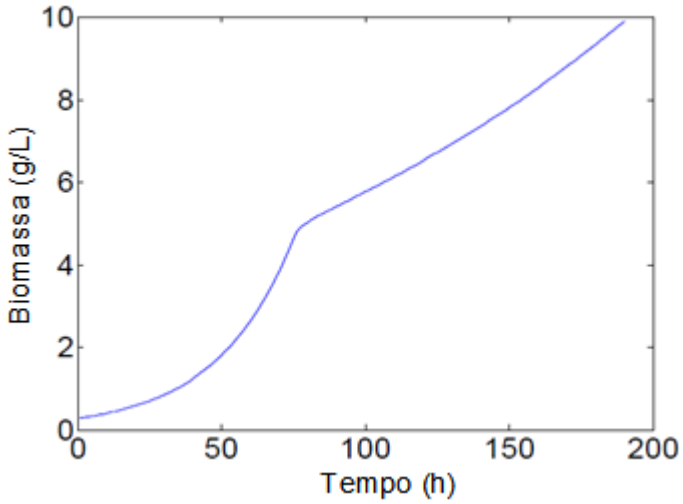

Figura 4.4: Modelo dinâmico para a concentração de biomassa no primeiro meio fermentativo. Fonte: Ranzan, C. (2014).

Em posse do conjunto de dados de entrada (espectros) e de saída (concentração de biomassa), é possível utilizar as versões do algoritmo ACO apresentadas no capítulo 3 na busca pelos pares de fluorescência que geram o melhor modelo preditivo.

A fim de verificar a contribuição de cada modificação para o resultado final, foram conduzidas cem replicações do algoritmo ACO com cada combinação de critérios (C1, C2) apresentada na Tabela 3.1. Cada replicação procurou pelo melhor modelo linear contendo três variáveis independentes ($\mathrm{k} = 3$). Este tamanho de modelo foi escolhido sem análise prévia a respeito do número ótimo de variáveis, sendo arbitrado apenas para permitir a construção e comparação de modelos pelos algoritmos, sem extender demasiadamente o tempo computacional. Como o objetivo deste trabalho não engloba a determinação do melhor tamanho de modelo para essa aplicação, justifica-se o uso de um número fixo arbitrário.

Cada replicação do algoritmo utilizou cinquenta ciclos de cem formigas e todos os modelos envolvidos foram obtidos pelo método de mínimos quadrados. Cabe salientar que, ao invés do número fixo de cinquenta ciclos, uma opção é utilizar a taxa de convergência como critério de parada. Isto evitaria que o algoritmo parasse antes de encontrar o mínimo, além de diminuir o tempo computacional, uma vez que algumas replicações encontram o ótimo logo nas primeiras iterações, mas continuam a percorrer todos os ciclos. Este aspecto, no entanto, foge ao escopo deste trabalho, sendo, portanto, sugerido como trabalho futuro.

Tendo-se sete critérios possíveis para atualização de trilha de feromônios e quatro critérios possíveis para seleção de modelo, existem 28 possíveis combinações de critérios a serem comparadas a fim de se encotrar a melhor delas. Para cada uma das 2.800 replicações (28 combinações x 100 replicações), foram computados o SSE, as variáveis selecionadas para o modelo e a evolução da trilha de feromônios. Ainda, uma busca exaustiva foi realizada para encontrar os três componentes que geram o melhor modelo, bem como o RMSEC mínimo possível. Essas informações serão importantes na quantificação da qualidade de outros modelos e permite a avaliação de cada versão (combinação de critérios C1 e C2) utilizada. O tempo computacional necessário para essa busca foi igual a 6223 segundos (1 hora e 43 minutos) , utilizando-se um processador Intel® Core™ i5 750 @ 2.67 GHz com 8 GB de mémoria RAM. A busca exaustiva consisitiu na avaliação do RMSEC de todos os modelos possíveis utilizando os dados da primeira fermentação, sendo que o menor valor encontrado (erro ótimo) foi igual a 0,19 g/L, associado ao modelo formado pelo pares espectrais 33, 49 e 50. Estes pares, portanto, provavelmente captam a fluorescência da espécie química NADPH, relacionada à respiração das leveduras, bem como a de outros fluoróforos biogênicos, como proteínas e coenzimas (Lindemann *et al.*, 1998; Pattison *et al.*, 2000). As regiões do espectro relacionadas a esses pares ótimos podem ser vistas na Figura 4.1.

4.1.2 Discussão dos resultados

Sabendo-se o menor erro possível para um modelo construído a partir do conjunto de dados em questão (fermentação 1), uma forma conveniente de comparar as 28 versões é analisar o número de vezes que cada uma encontrou o resultado ótimo (RMSEC mínimo) dentre todas as 100 replicações. Essa comparação é apresentada na Figura 4.5.

Lembrando que as combinações 1 a 16 utilizam métricas globais e as combinações 17 a 28 utilizam métricas individuais para a atualização da trilha de feromônios, pode-se perceber na Figura 4.5 uma leve vantagem desse segundo grupo, destacando-se as combinações 17 (melhor caso), 18, 19, 20 e 27.

No entanto, sabe-se que cada uma das versões utilizadas é formada por dois critérios e, portanto, convém avaliar separadamente a influência dos mesmos no resultado encontrado. Para isso considerou-se, do total de replicações que encontraram o erro ótimo, a fração atribuída à utilização de cada critério C1 e cada critério C2.

Para analisar a influência do critério C1, foram ignoradas as 4 variações de C2, ou seja, as 28 versões do algoritmo foram agrupados de 4 em 4, gerando 7 grupos que utilizam determinado C1. Por exemplo, as versões 1 a 4, que utilizam o SSE como C1 (vide Tabela

3.1) tiveram suas frações somadas e o resultado foi atribuído a esse critério de atualização de trilhas. O mesmo foi feito para os outros 6 critérios C1, e o resultado pode ser visto na Figura 4.6.

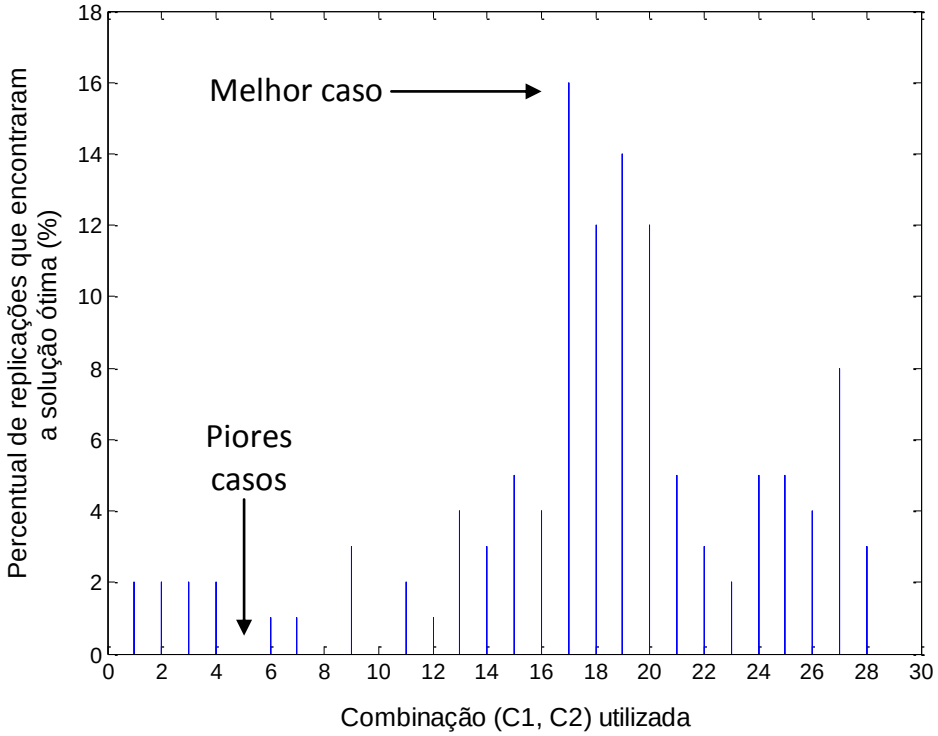

Figura 4.5: Percentual de vezes que cada combinação de critérios encontrou o resultado ótimo, definido pela busca exaustiva, dentre todas as 100 replicações.

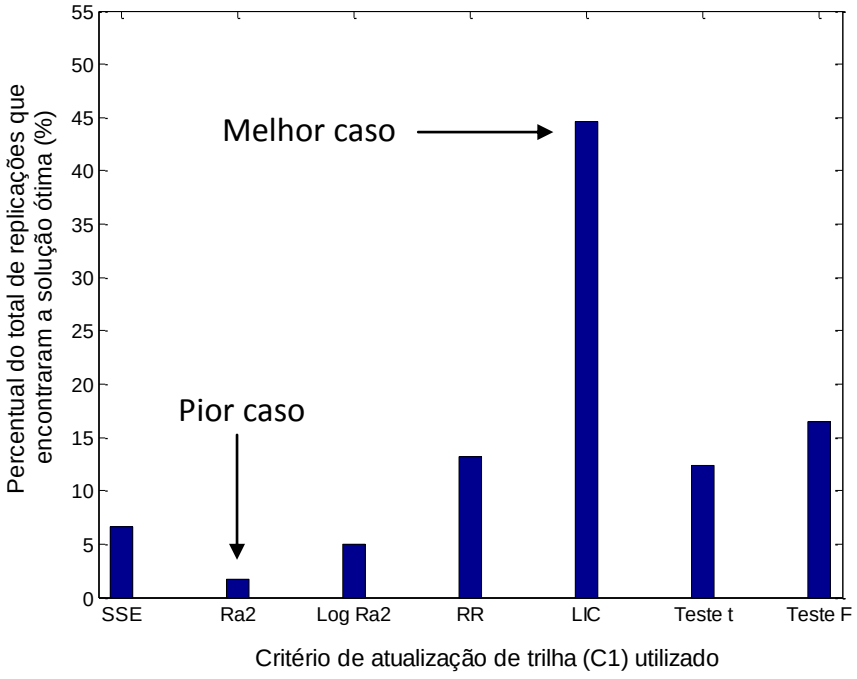

Figura 4.6: Influência de C1: percentual do total de replicações que atingiram a solução (erro) ótima associado às combinações de critérios que utilizam o mesmo critério de atualização de trilha.

Conforme visto na Figura 4.6, a utilização do tamanho do intervalo de confiança (LCI) como C1 gerou melhores resultados, encontrando o modelo ótimo com mais frequência que os outros casos. De todas as 120 replicações que encontraram o ótimo, distribuídas ao longo de 28 combinações, 45% pertencem ao grupo que utiliza o LCI para atualização da trilha. Esse é um número bastante significativo frente aos resultados encontrados pelo modelo ACO original, representado pela combinação de critérios número 1, que responde por apenas 7% das replicações que encontraram o ótimo. Desta forma, o uso do valor do LCI é um significativo aprimoramento na busca pelas variáveis mais representativas.

Para analisar a influência do critério C2, foram ignoradas as variações de C1, ou seja, as combinações que utilizam o mesmo critério C2 foram agrupados, gerando 4 grupos. Por exemplo, as combinações 1, 5, 9, 13, 17, 21 e 25, que utilizam o SSE como C2 (vide Tabela 3.1) tiveram suas frações somadas e o resultado foi atribuído a esse critério de seleção de modelos. O mesmo foi feito para os outros 3 critérios C2, e o resultado pode ser visto na Figura 4.7.

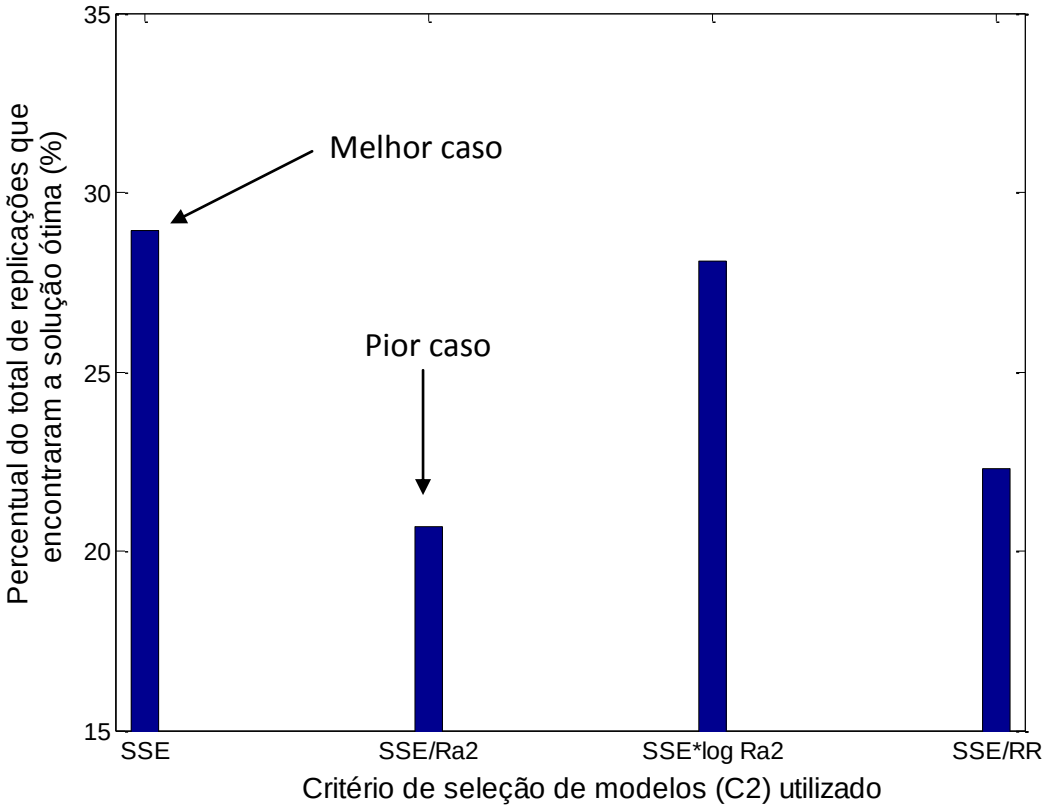

Figura 4.7: Influência de C2: percentual do total de replicações que encontraram a solução ótima associado às combinações de critérios que utilizam o mesmo critério de seleção de modelos.

A Figura 4.7 mostra que o pior caso é o uso da razão SSE/R_a^2 como C2, responsável por 20% das replicações que encontraram o erro ótimo. No entanto, quando se observa o melhor caso (29% atribuídos ao uso do SSE), percebe-se que há uma diferença de apenas 9% entre os dois extremos. Assim, embora a utilização do SSE como C2 mostre uma leve vantagem, as diferentes métricas para seleção de modelos (C2) parece não ser

significativa para o resultado final, uma vez que os quatro grupos geraram frações semelhantes.

Além da qualidade de otimização das 28 combinações de critérios, também é necessário avaliar a capacidade de geração de bons modelos preditivos. Para isso, os 100 modelos obtidos por cada uma das 28 combinações na fase de calibração foram aplicados aos dados da segunda fermentação (conjunto de dados de teste), e seus respectivos RMSEP foram analisados. A Figura 4.8 mostra o valor de RMSEP capaz de abranger o resultado de 90% das replicações realizadas por cada combinação de critérios, ou seja, o 90º percentil dos erros encontrados. Por exemplo, usando-se a versão 1, 90% de todas as replicações encontraram um erro menor que 0,48 g/L.

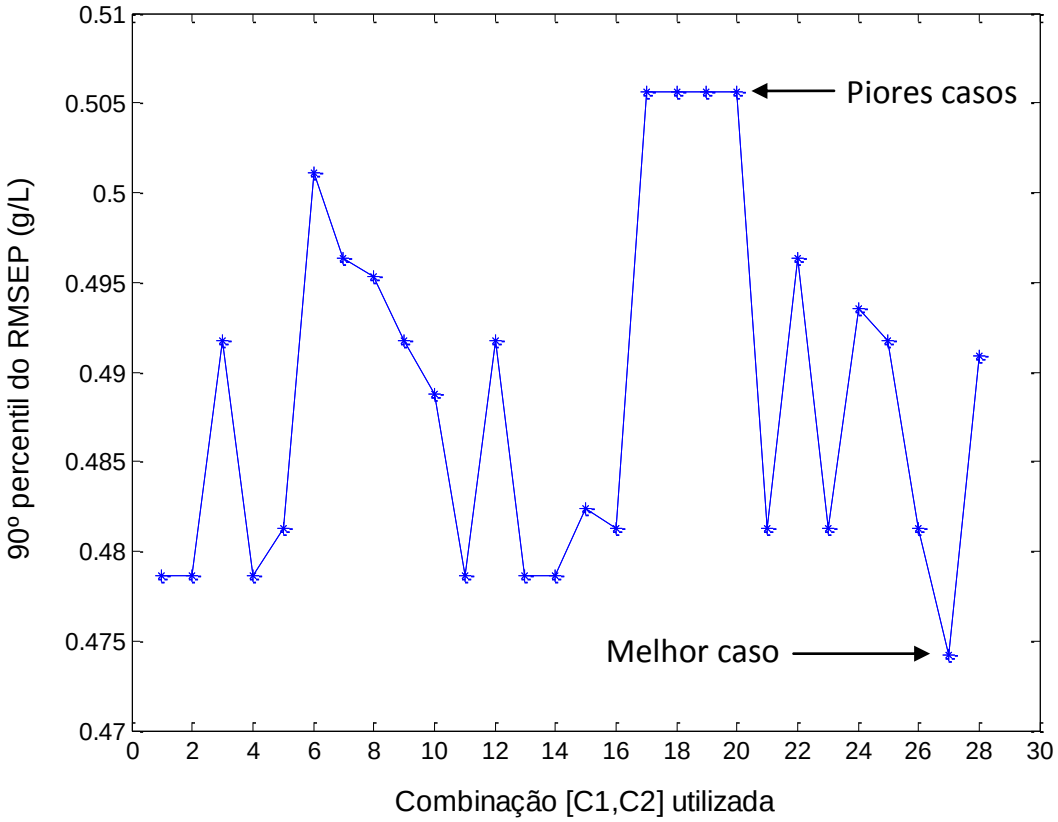

Figura 4.8: Valor do RMSEP capaz de abranger 90% dos erros encontrados nas replicações realizadas por cada combinação de critérios (90º percentil), utilizando os dados de teste.

A Figura 4.8 mostra que, embora as combinações 17 a 20 tenham fornecidos os melhores resultados na fase da calibração, os mesmos apresentam os maiores erros quando aplicados a um novo conjunto de dados. Enquanto a versão 1 apresenta 90% dos modelos com erros menores que 0,48 g/L, esse valor sobe para 0,51 no caso dos grupos com C1 igual ao LCI. O melhor resultado no que diz respeito à predição foi o par de critérios 27, com um 90º percentil igual 0,475 g/L. No entanto, quando se analisa a diferença entre os valores mínimo e máximo do 90º percentil, percebe-se que os mesmo diferem por apenas 0,03 g/L, o que é pequeno frente à ordem de grandeza dos dados de saída do modelo.

Diante da evidência de bons resultados encontrados na fase de calibração pelas versões do algoritmo que utilizam o tamanho do intervalo de confiança como C1, em

especial a combinação 17, convém analisar mais a fundo sua capacidade preditiva. Para isso, os dois modelos descritos abaixo foram utilizados para predizer os valores de concentração de biomassa da fermentação 2 (grupo de amostras de teste). Essas predições foram então plotadas contra os dados fornecidos pelo modelo dinâmico da mesma fermentação, e o resultado é apresentado na Figura 4.9.

1) Modelo construído utilizando-se os pares espectrais [33, 50, 57], os quais forneceram o modelo ótimo durante a fase de calibração, com um RMSEC mínimo igual a 0.19 g/L.

2) Modelo construído utilizando-se os pares espectrais [33, 49, 50], os quais foram os componentes espectrais escolhidos mais frequentemente na fase de calibração pela versão do algoritmo ACO que utiliza o tamanho do intervalo de confiança como C1 e o SSE como C2 (combinação 17).

Os números dentro dos colchetes referem-se a um par de emissão/excitação analisado na espectroscopia de fluorescência. Para verificar quais comprimentos de onda são utilizados nesses pares, vide Figura 4.1.

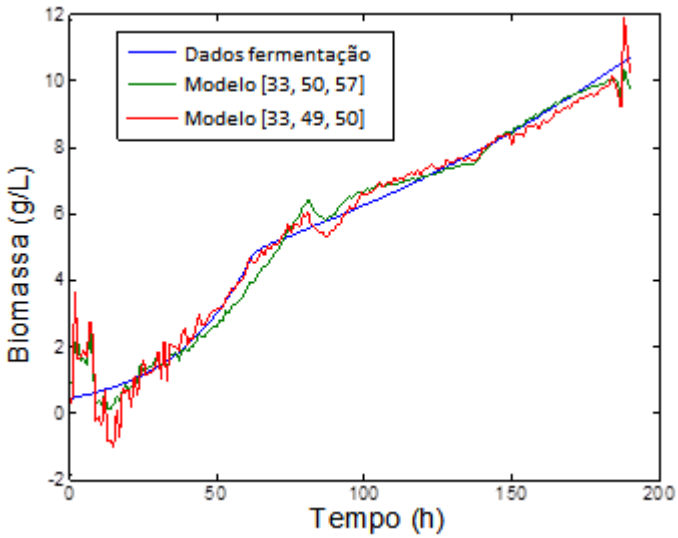

Figura 4.9: Comparação entre os dados previstos pelo modelo dinâmico (azul), pelo melhor modelo quimiométrico encontrado na fase de calibração (verde) e pelo modelo formados pelos componentes escolhidos com mais frequencia pelo versão 17 do algoritmo ACO (vermelho).

Conforme mostrado na Figura 4.9, embora a maioria das replicações não tenha encontrado a solução ótima global, o modelo formado pelos componentes encontrados mais frequentemente pela versão 17 do ACO é muito semelhante a ela. Ambos os modelos (o ótimo e o mais frequente) foram capazes de predizer a concentração de biomassa no meio com um RMSEP razoável: 0,42 e 0,54 g/l, respectivamente. Embora os dois modelos não sejam exatamente os mesmos, dois dos componentes são iguais (33, 50). A diferença de predição deve-se então ao terceiro componente de cada modelo (57 e 49), os quais residem em regiões espectrais distintas.

Apesar do valor de RMSEP encontrado ser satisfatório, quando se analisa os valores gerados pelos outros pares de critérios, obtidos a partir do modelo formado pelos componentes mais frequentes e apresentados na Figura 4.10, percebe-se que, na verdade, o par 17 encontrou o pior valor de RMSEP. Da mesma forma, os pares 18 a 20, que também utilizam o LCI, encontraram esse mesmo valor.

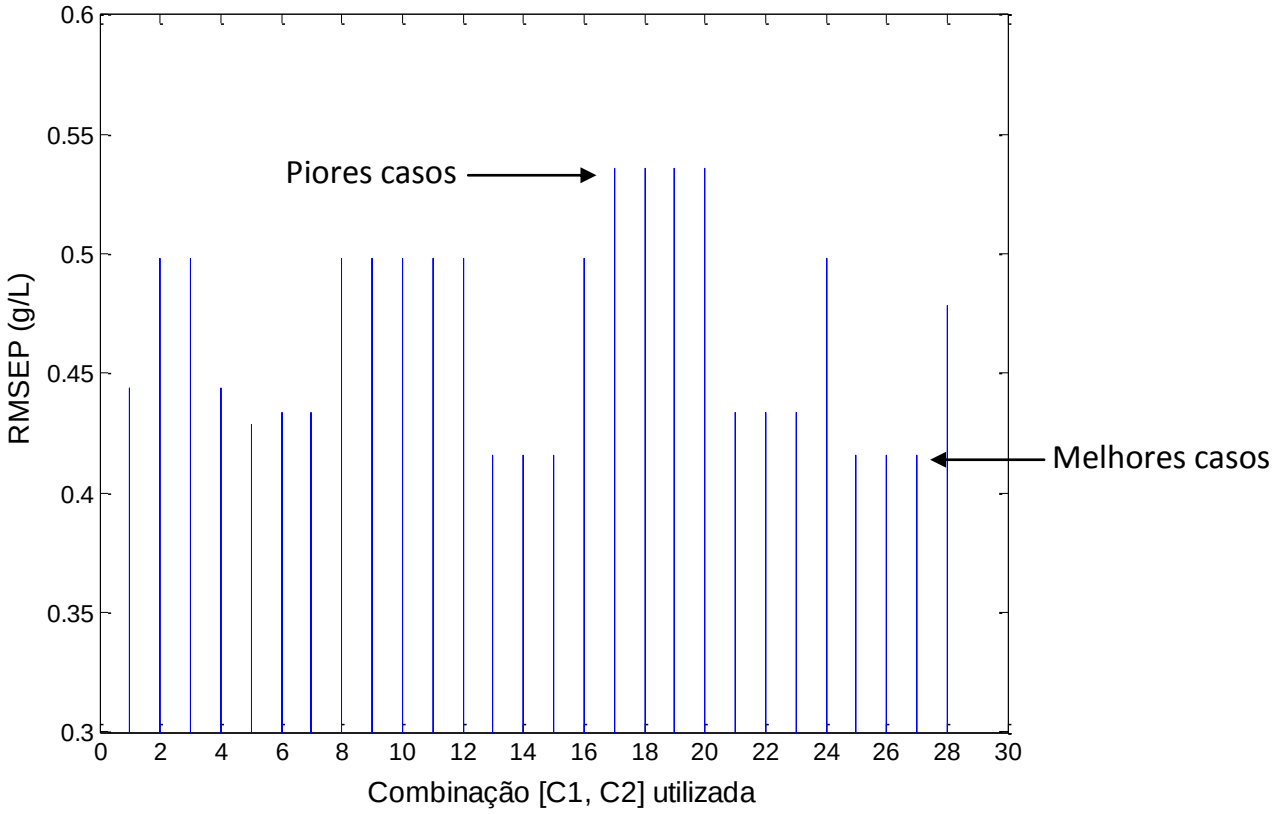

Figura 4.10: Valor do RMSEP encontrado pelo modelo formado pelos componentes mais frequentes de cada par de critérios quando aplicado ao conjunto de dados de teste (2ª fermentação).

Assim, dentre os diferentes critérios propostos para atualização da trilha, o tamanho do intervalo de confiança (LCI) foi o que apresentou o melhor desempenho de otimização, independendo do critério utilizado na seleção de modelos. O grupo que utilizou o LCI (pares 17, 18, 19 e 20) encontrou a solução ótima em 13,5% das suas replicações e foi responsável por 45% de todas as vezes que o modelo ótimo foi encontrado. Além disso, os componentes selecionados com maior frequência pela versão 17 formam um modelo próximo ao modelo ótimo, do qual difere por apenas um componente espectral. No entanto, quando comparado às outras versões, a qualidade preditiva do modelo construído é inferior, gerando um RMSEP 29% maior que o RMSEP mínimo encontrado na fase teste. Nesse último quesito, as combinações 13 – 15 (RR como C1) e 25 - 27 (Teste F como C1) obtiveram o melhor resultado, pois os componentes que os mesmos escolheram com mais frequência são aqueles que formam o modelo ótimo, com um RMSEP igual a 0,42 g/L.

Outra conclusão interessante é que 3 das 4 versões que utilizam o RR e o Teste F como C1 foram capazes de selecionar com maior frequência os componentes que formam o modelo ótimo, e 3 das 4 versões que utilizam o Teste t obtiveram um modelo

com RMSEP muito próximo ao ótimo. Isso sugere que o uso desses três critérios de atualização de trilha geram modelos com uma melhor capacidade preditiva. Além disso, com exceção da versão 4, o uso do critério SSE/RR como C2 parece prejudicar o resultado, pois, para um mesmo C1, as versões que o utilizaram geraram RMSEP maiores que os que utilizaram outros critérios de seleção de modelos.

4.2 Conteúdo protéico da farinha

4.2.1 Descrição

No estudo de caso anterior, as modificações sugeridas para o algoritmo foram testadas utilizando-se uma matriz reduzida de dados espectrais, contendo apenas 150 pares de fluorescência. Isso permitiu verificar mais facilmente as diferenças entre as versões propostas, bem como identificar que a alteração dos critérios de seleção de modelo não modifica significativamente o resultado. Diante disso, convém aprofundar os estudos utilizando-se uma matriz de dados mais complexa que contenha um número maior de possíveis preditores. O presente estudo de caso tem essa tarefa como objetivo, utilizando para isso dados de espectroscopia NIR, amostras de diferentes marcas de farinha e o algoritmo ACO sugerido anteriormente como ferramenta de seleção dos melhores comprimentos de onda. As modificações implementadas na atualização da trilha de feromônios do ACO são as mesmas que aquelas utilizadas no estudo de caso de fermentação.

Os dados experimentais utilizados no presente estudo incluem 34 amostras de diferentes marcas e tipos de farinha, medidas em triplicata para a determinação de seus conteúdos proteicos, em g/kg. Foram feitas, portanto, 102 análises, as quais foram divididas de forma alternada para a formação dos conjuntos de calibração e teste. A Figura 4.11 mostra que, da forma como foram divididos, esses dois conjuntos de dados abrangem as mesmas regiões de informação, garantindo, assim, uma varredura de todo o espaço amostral, tanto na calibração quanto no teste do modelo.

Todas as medições de conteúdo proteico foram feitas off-line, utilizando-se um Aparelho de Digestão (Digesdahl® Hach - Düsseldorf, Alemanha). Paralelamente a essas análises farinográficas, as amostras foram caracterizadas através de medições de espectroscopia NIR em um Analisador NIR Multifuncional (Bruker Optik GmbH - Ettlingen, Alemanha), variando-se o comprimento de onda de 800 nm até 2800 nm. O intervalo entre os comprimentos de onda é variável, de forma que são analisados 1150 elementos espectrais no total. Todas as implementação e cálculos envolvidos foram feitos pelo software MATLAB (Ver. 5.3.0.10183 R11, The Mathworks, Inc., Natick, USA) (Ranzan, C., 2014).

Foi arbritado um tamanho de modelo igual ao do caso anterior ($k = 3$), a fim de permitir futuras comparações entre os dois estudos. Cada uma das 28 versões realizou 100 vezes a busca pelo ótimo, apresentando, em cada replicação, o valor RMSEC do melhor modelo encontrado e os três componentes espectrais que o geraram. Cada busca utilizou cinquenta ciclos de cem formigas no algoritmo e todos os modelos envolvidos foram obtidos pelo método de mínimos quadrados. Devido à extensão do conjunto de dados, a busca exaustiva não foi realizada.

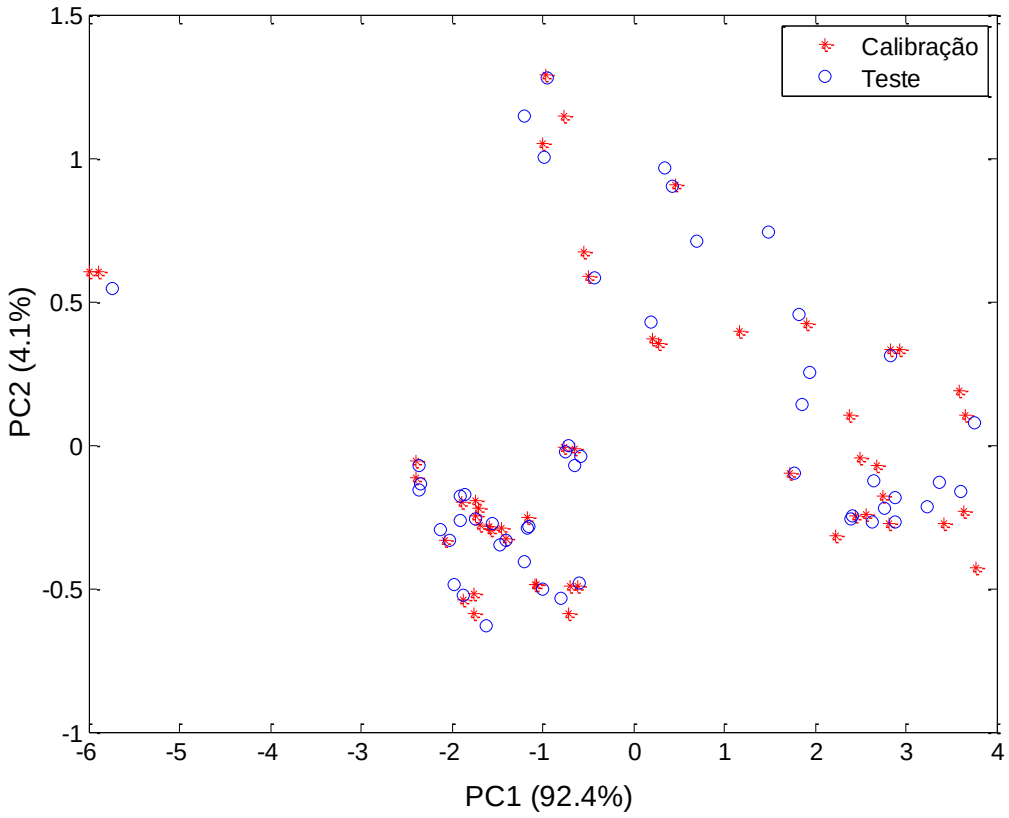

Figura 4.11: Componente principal 1 versus componente principal 2 para os dois conjuntos de amostras de farinha (calibração em vermelho e teste em azul).

4.2.2 Discussão de resultados

Diferentemente do estudo de caso anterior em que se realizou a busca exaustiva, nesse estudo não é conhecida a combinação de componentes que gera o melhor modelo, nem o seu respectivo RMSEC. Dessa forma, a comparação entre as versões não pode ser feita utilizando-se a frequência com que o modelo ótimo foi encontrado. Porém, verificou-se que de todas as 2800 replicações (28 critérios x 100 replicações), o menor valor encontrado para o erro foi de 0,55 g/kg e esse, portanto, será o erro mínimo utilizado como referência. Assim, verificou-se o percentual de replicações que cada combinação de critérios encontrou esse erro mínimo, e o resultado pode ser visto na Figura 4.12.

A Figura 4.12 sugere um resultado bem diferente daquele obtido no estudo de caso anterior. As versões que utilizam o tamanho do intervalo de confiança (17 a 20) não encontraram erros iguais ou próximos ao erro mínimo. Os melhores resultados, nesse caso, são obtidos pelas versões 21, 23 e 24. Essas três versõess tem em comum o uso do Teste t como critério de atualização de trilha.

No entanto, sabe-se que cada uma das combinações utilizadas é formada por dois critérios e, portanto, convém avaliar separadamente a influência dos mesmos no resultado encontrado. Para isso considerou-se, do total de replicações que encontraram o erro ótimo, a fração atribuída à utilização de cada critério C1 e cada critério C2.

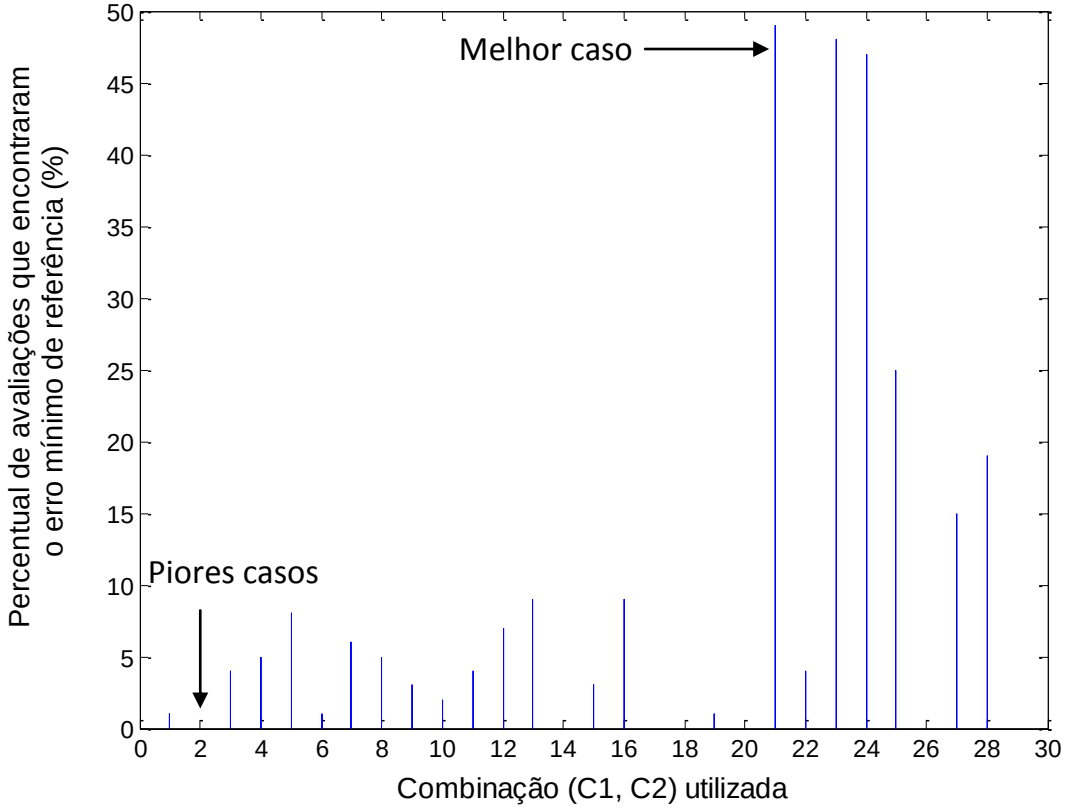

Figura 4.12: Percentual de replicações que cada combinação de critérios obteve um modelo com RMSEC igual ao erro mínimo de referência, dentre todas as 100 replicações.

Para analisar a influência do critério C1, foram somadas as 4 variações de C2, ou seja, as 28 versões foram agrupados de 4 em 4, gerando 7 grupos que utilizam o mesmo C1. Cada um dos 7 grupos formados é responsável por uma fração do total de replicações que obtiveram erros mínimos. Esse resultado pode ser visto na Figura 4.13.

Conforme visto na Figura 4.13, a utilização do Teste t como C1 gerou melhores resultados, encontrando os modelos com erros mínimos mais frequentemente que os outros casos. De todas as 263 replicações que encontraram um erro igual ao erro mínimo, distribuídas ao longo de 28 versões, 53% pertencem ao grupo que utiliza o Teste t para atualização da trilha. Esse é um número bastante significativo frente aos resultados encontrados pelo modelo ACO original, representado pelo par de critérios número 1, que responde por apenas 4% dessas replicações. Desta forma, para este tipo de dados, o uso da estatística t pode ser um importante aliado na busca pelas variáveis mais representativas.

Para analisar a influência do critério C2, foram somadas as variações de C1, ou seja, as versões que utilizam o mesmo critério C2 foram agrupados. Cada um dos 4 grupos formados é responsável por uma fração do total de replicações que obtiveram erros mínimos . Esse resultado pode ser visto na Figura 4.14.

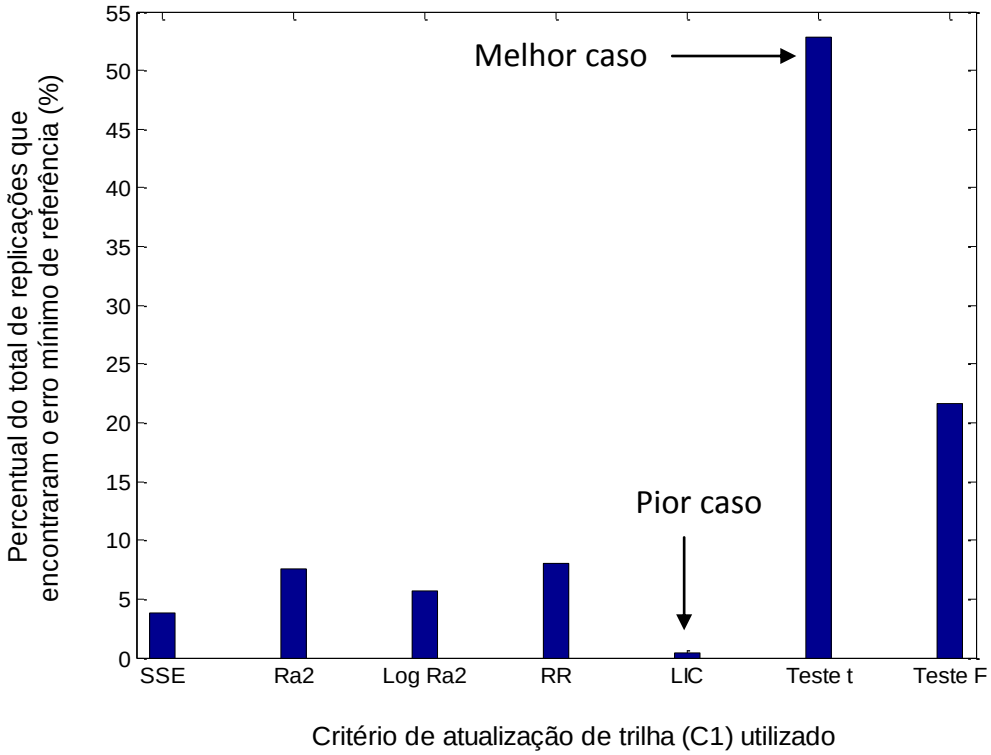

Figura 4.13: Influência de C1: percentual do total de replicações com erros iguais ao erro mínimo de referência associado às versões do ACO que utilizam o mesmo critério de atualização de trilha.

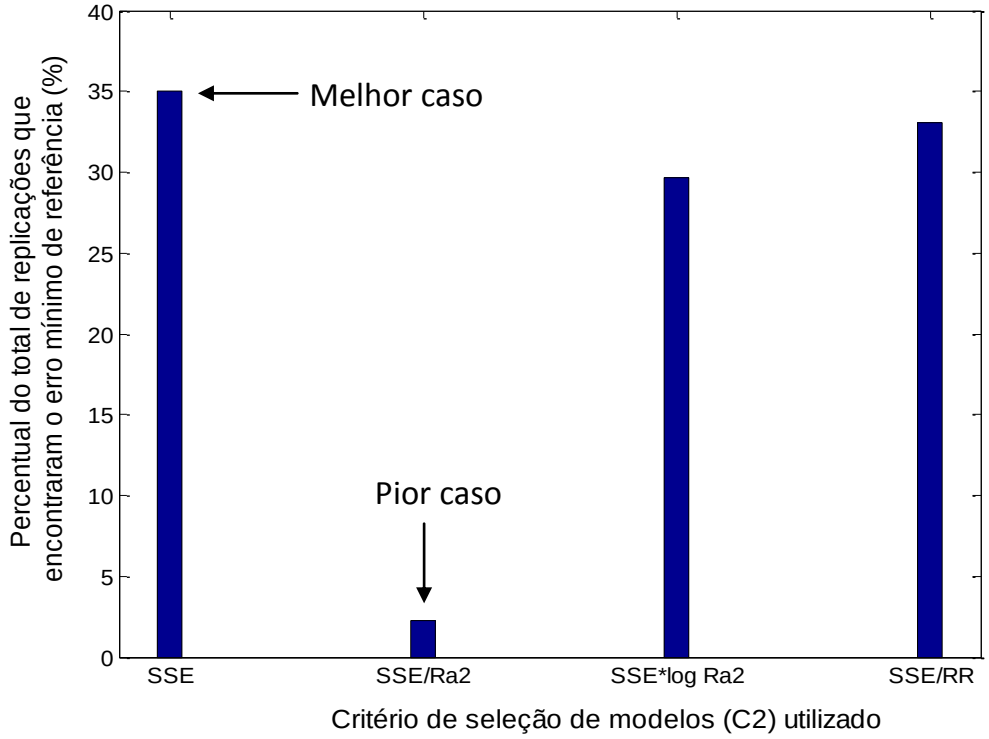

Figura 4.14: Influência de C2: percentual do total de replicações com erros iguais ao erro mínimo encontrado associado à versões do ACO que utilizam o mesmo critério de seleção de modelos.

A Figura 4.14 mostra que o pior caso é o uso da razão SSE/R_a^2 como C2, responsável por apenas 2% (ou 0,02) das replicações que encontraram os erros mínimos. Esse resultado vai ao encontro daquele obtido no estudo de caso anterior, onde esse critério também forneceu o menor percentual. No entanto, quando se observa os outros casos, percebe-se que não há uma diferença significativa entre eles, sendo cada um responsável por cerca de 30% das replicações. Assim, com exceção da utilização da razão SSE/R_a^2 como critério C2, que fornece o pior resultado, as diferentes métricas para seleção de modelos parece não ser significante para o resultado final, uma vez que os outros três grupos possuem frações semelhantes.

Tendo avaliado a capacidade de otimização de cada versão do ACO, convém avaliar a capacidade de predição dos modelos gerados por eles. Para isso, os 100 modelos de cada um das 28 versões foram aplicados ao conjunto de dados de teste, e seus respectivos erros computados. A partir disso, foi calculado o 90º percentil dos erros obtidos, dado pela

Figura 4.15, a fim de observar o limiar que abrange a grande maioria das replicações. Além disso, foi avaliada a capacidade preditiva dos 28 modelos, um para cada versão, formados pelos componentes espectrais escolhidos mais frequentemente. Os valores RMSEP encontrados por cada modelo podem ser vistos na Figura 4.16.

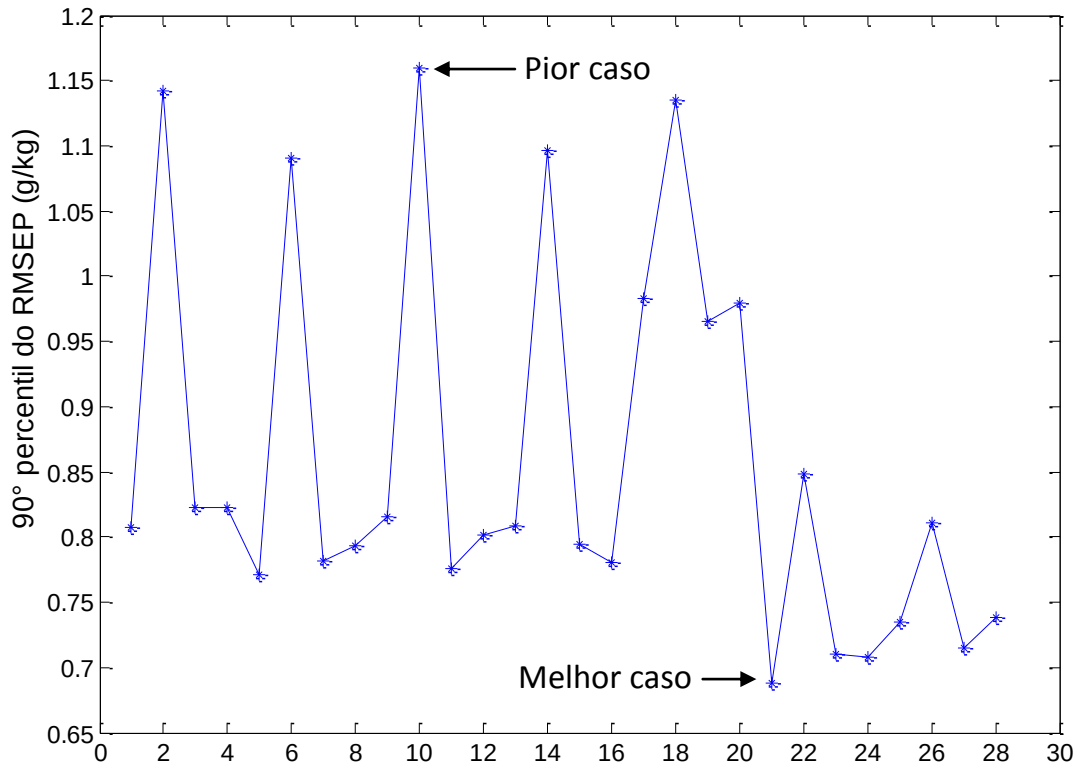

Figura 4.15: Valor RMSEP capaz de abranger 90% dos erros encontrados nas replicações realizadas por cada combinação de critérios (90º percentil), utilizando os dados de teste.

Quando aplicado a um novo conjunto de dados, o melhor modelo da fase de calibração fornece um erro de 0,64 g/kg, considerado aqui o erro mínimo de teste. Com base nesta informação e pela Figura 4.15, conclui-se que os melhores resultados são referentes às versões que utilizam o Teste t ou o Teste F para atualização da trilha (21, 23, 24 e 25, 27, 28), pois 90% das replicações geraram erros menores que aproximadamente 0,7 g/kg, valor muito próximo ao erro mínimo. É interessante notar na Figura 4.15 a evidente influência negativa da utilização do critério SSE/R_a^2 como C2, implícita nas versões 2, 6, 10, 14, 18, 22 e 26.

A Figura 4.16, por sua vez, mostra o valor do RMSEP encontrado pelos modelos formados pelos componentes selecionados com mais frequência por cada combinação de critérios. Assim, percebe-se que o modelo obtido pela versão 23 foi o que obteve o menor RMSEP (0,6 g/kg) quando aplicado aos dados de teste. No entanto, pode-se observar na Figura 4.16 que os erros estão bastante distribuídos, não se destacando nenhum tipo de padrão.

Considerando que a média dos conteúdos proteicos medidos é 9,5 g/kg, pode-se afirmar que a maioria das versões fornecem modelos com um erro razoável de predição, pois os valores de RMSEP encontrados concentram-se na região entre 0,8 g/kg e 1,5 g/kg, equivalente a um erro de 8% e 16%, respectivamente.

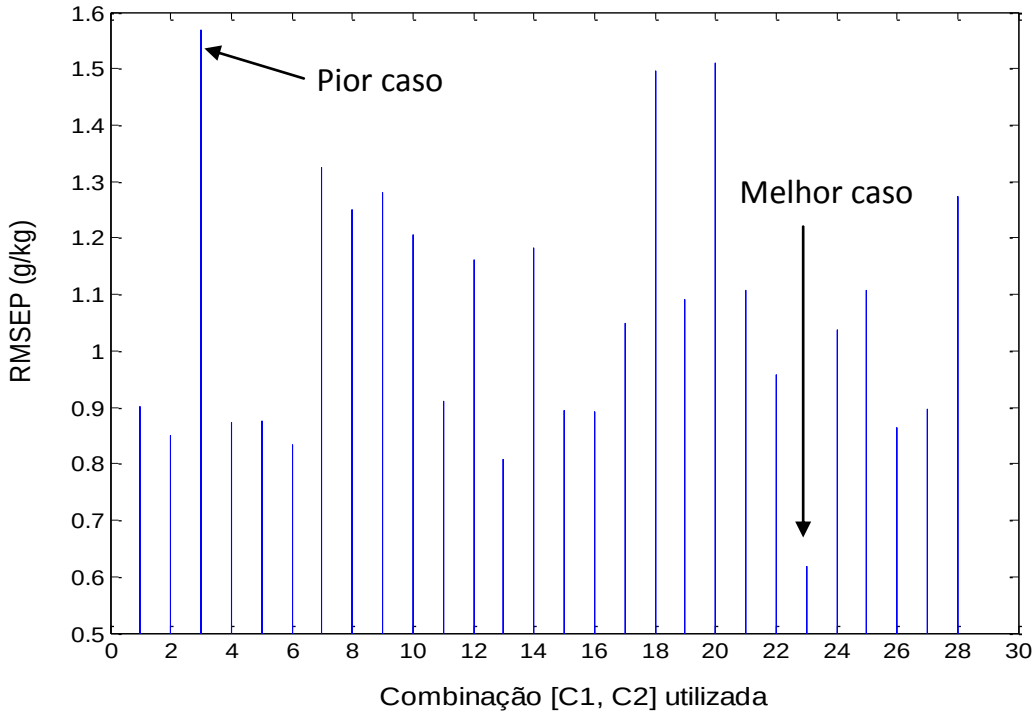

Figura 4.16: Valor do RMSEP obtido pelo modelo formado pelos componentes mais frequentemente encontrados por cada combinação de critérios quando aplicado ao conjunto de dados de teste.

Assim, os resultados mostram que a utilização do Teste t na atualização da trilha de feromônios forneceu o melhor desempenho de otimização, uma vez que foi responsável por 53% das replicações com erros mínimos. O segundo melhor critério C1 foi Teste F, responsável por 22% dessas replicações. Assim, da mesma forma que no estudo anterior, os melhores resultados foram obtidos pela utilização de métricas de desempenho

individual dos componentes como critério de atualização de trilha, ainda que os dois experimentos tenham destacado critérios distintos.

No que diz respeito à capacidade de predição dos modelos obtidos, a maioria das versões parecem se equivaler quanto à qualidade preditiva, excetuando-se aqueles que encontraram um RMSEP próximo a 1,5 g/kg e a versão 23, com uma capacidade preditiva levemente superior.

4.3 Conclusões

A realização dos dois estudos de caso buscou permitir a análise do comportamento das versões sugeridas quando submetidas a matrizes de dados provenientes de duas técnicas espectrométricas diferentes e com diferentes complexidades. Assim, o experimento fermentativo consistia de 150 possíveis preditores, referentes aos pares de espectroscopia de fluorescência 2D; o segundo experimento, por sua vez, consistia de 1150 componentes espectrais provenientes de espectroscopia NIR.

Enquanto o estudo da fermentação indicou que a utilização do tamanho do intervalo de confiança como C1 forneceu os melhores resultados, o estudo utilizando amostras de farinha sugere que o melhor critério C1 é o Teste t. Apesar da diferença conceitual, ambos os critérios utilizam o princípio de quantificação individual dos componentes do modelo, que parece resultar num melhor método de otimização.

No que diz respeito à modificação dos critérios de seleção de modelos, os dois estudos mostram que, com exceção do critério SSE/R_a^2, que fornece resultados significativamente piores, não há distinção na utilização dos diferentes critérios C2.

No entanto, quando se analisa a qualidade dos versões para geração de modelos preditivos, os dois estudos de caso apresentaram resultados bastante diferentes. No primeiro estudo de caso, a melhor versão na fase de calibração gerou o pior RMSEP na fase de teste. Mesmo assim, o valor encontrado (0,54 g/L) ainda é satisfatório diante da ordem de grandeza dos dados de saída (concentração de biomassa). No segundo estudo de caso, as 28 versões parecem se equivaler quanto à qualidade de predição, ainda que a versão 23 tenha gerado um modelo com RMSEP mínimo de 0,6 g/kg e algumas versões tenham resultado em um RMSEP em torno de 0,9 g/kg. De um modo geral, a capacidade preditiva dos modelos no segundo estudo de caso foi inferior às do primeiro estudo, o que provavelmente se deve à diferença na extensão dos dados.

A fim de determinar a melhor versão, convém avaliar de forma mais objetiva o desempenho de cada um deles nas análises dos estudos de caso. Esse desempenho pode ser atribuído a uma nota, que corresponde à posição que determinada versão obteve no ordenamento dos resultados. Como foi dito anteriormente, em cada estudo foi avaliado três aspectos: capacidade de otimização, reprodutibilidade de baixos erros e qualidade de predição do modelo mais frequente. Assim, a partir da comparação entre as versões, cada um recebeu seis notas: três referentes ao primeiro estudo de caso e três referentes ao segundo estudo de caso. Por exemplo, na Figura 4.5 pode-se ver que a versão 17 apresentou o melhor resultado e, portanto, recebe a nota 1 nesta análise. Desta forma, a melhor versão será aquela que obtiver a menor soma das seis notas.

Assim, conforme pode ser visto na Tabela 4.1, as versões que apresentaram o melhor desempenho geral são a versão 23, com uma nota igual a 20, e a versão 27, com uma nota igual a 26. Por outro lado, a versão 18 corresponde ao pior resultado, totalizando 90 pontos. Ainda, convém ressaltar que a maioria das modificações implementadas superaram o ACO original (versão 1), que obteve nota igual a 54. Isso se deve, sobretudo, ao segundo estudo de caso, uma vez que a versão 1 obteve bons resultados no primeiro estudo, especialmente na análise de reprodutibilidade e de predição .

Tabela 4.1: Quadro de notas das versões do ACO para as 6 análises feitas ao longo dos dois estudos de caso.

	Fermentação			Farinha			
Método	Otimização	Reprodutibilidade	Predição	Otimização	Reprodutibilidade	Predição	SOMA
1	8	2	4	15	14	11	54
2	8	2	6	16	26	4	62
3	8	7	6	12	18	28	79
4	8	2	4	11	18	6	49
5	10	3	2	8	7	7	37
6	9	11	3	15	23	3	64
7	9	10	3	10	10	25	67
8	10	9	6	11	11	22	69
9	7	7	6	13	17	24	74
10	10	5	6	14	27	21	83
11	8	2	6	13	8	12	49
12	9	7	6	9	13	19	63
13	6	2	1	7	15	2	33
14	7	2	1	16	24	20	70
15	5	4	1	13	12	9	44
16	6	3	6	7	9	8	39
17	1	12	7	16	22	15	73
18	3	12	8	16	25	26	90
19	2	12	7	15	20	16	72
20	3	12	7	16	21	27	86
21	5	3	3	1	1	18	31
22	7	10	3	13	19	13	65
23	8	3	3	2	3	1	20
24	5	8	6	3	2	14	38
25	5	7	1	4	5	17	39
26	6	3	1	16	16	5	47
27	4	1	1	6	4	10	26
28	7	6	5	5	6	23	52

Capítulo 5 – Caracterização do Diesel combinando técnicas espectrométricas – Avaliação Preliminar

Neste capítulo, as diferentes versões do algoritmo ACO são utilizadas para criar modelos de predição do conteúdo de enxofre em amostras de diesel, utilizando como entrada os dados obtidos por dois tipos de espectroscopia. Com isso, busca-se, sobretudo, avaliar a aplicabilidade das versões propostas no ramo de caracterização de combustíveis, bem como identificar aquele que apresenta melhor desempenho.

Inicialmente, as amostras de diesel utilizadas são detalhadas e as condições em que as duas técnicas espectrométricas foram conduzidas são especificadas. Em seguida apresenta-se a forma como foram criados os modelos de predição de enxofre e a discussão dos resultados obtidos por cada versão ACO.

5.1 Apresentação de amostras

O objetivo deste capítulo consiste em estudar o desempenho das versões do ACO modificado e apresentado anteriormente quando aplicado na caracterização de combustíveis, especificamente na determinação do conteúdo de enxofre em diesel, utilizando-se duas técnicas espectrométricas diferentes concomitantemente, a espectroscopia de fluorescência bidimensional e a espectroscopia NIR. Para isso, foram utilizadas 45 amostras de combustível diesel S100 provenientes diretamente da corrente de saída de hidrotratamento (diesel HDT) da Refinaria Alberto Pasqualini (REFAP) e dos tanques de armazenagem do mesmo. As amostras foram coletadas por pessoal especializado e seguiram o mesmo processo de análise certificadora utilizada como rotina dentro da refinaria.

A coleta dos pontos aconteceu uma vez ao dia em um período espaçado de três meses. Assim, as amostras de diesel são oriundas de diferentes cargas de petróleo, o que resulta em amostras com prováveis diferenças em constituição final.

O teste utilizado para certificação de concentração de enxofre das amostras foi o ASTM D-7039, teste padrão regulamentado pela ANP para quantificar concentração de enxofre para diesel S500 e S10. Cada amostra foi medida em triplicata, totalizando 135 medições, que foram separadas de forma alternada em dois grupos: um grupo para estimação de parâmetro, contendo 68 medições, e outro para teste, com 67 medições. Da forma como esses dados foram divididos, além de facilitar os cálculos e diminuir o tempo computacional em relação a, por exemplo, uma validação cruzada, esses dois conjuntos abrangem as mesmas regiões de informação, conforme mostra a Figura 5.1 para os espectros de flurorescência e a Figura 5.10 para os espectros de NIR. Isso garante uma varredura de todo o espaço amostral, tanto na calibração quanto no teste do modelo.

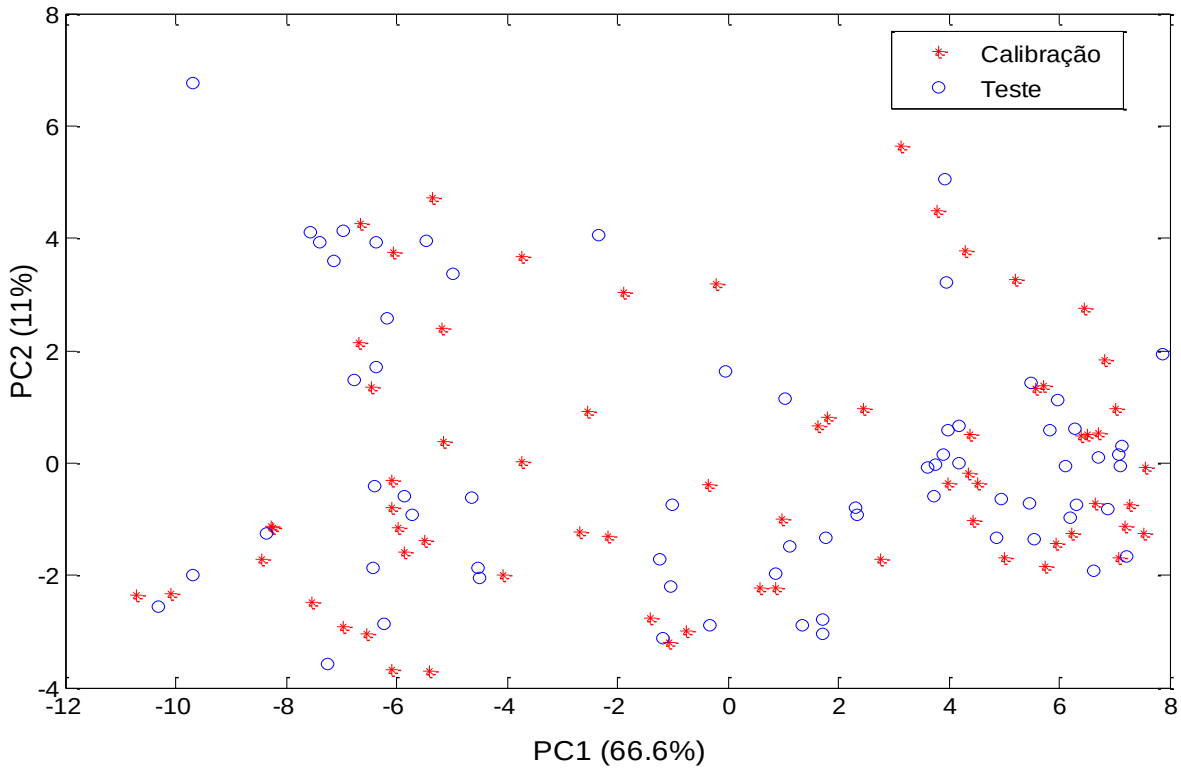

Figura 5.1: Componente principal 1 versus componente principal 2, para os dois conjuntos de dados (calibração teste) de fluorescência bidimensional.

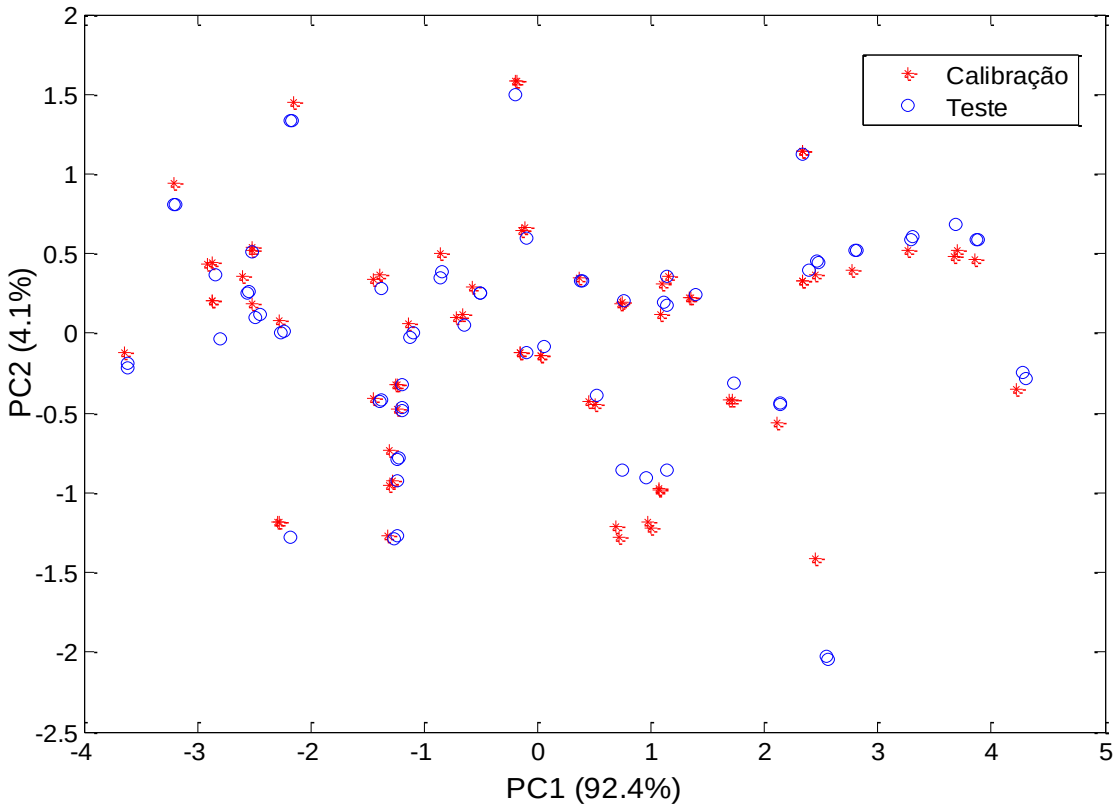

Figura 5.2: Componente principal 1 versus componente principal 2, para os dois conjuntos de dados (calibração teste) de infravermelho próximo.

5.2 Espectroscopia de Fluorescência 2D

A primeira técnica espectrométrica utilizada consiste na espectroscopia de fluorescência bidimentsional. Os espectros de fluorescência das amostras diesel foram medidos com uso do equipamento HORIBA Fluoromax®-4, equipado com lâmpada de xenônio de 150 W e range de emissão e excitação entre 200 nm – 950 nm. O equipamento consiste basicamente de uma fonte de excitação, dois monocromadores para seleção de comprimento de onda, um de excitação e outro de emissão e módulo de amostragem com detector de referência e um detector de emissão. A coleta foi realizada com o uso de fibra ótica, onde as amostras foram acondicionadas em frascos de vidro e inseridas em uma câmara escura com entrada para a fibra ótica, ilustrada na Figura 5.3. O uso desta metodologia com fibra ótica e câmara escura foi proposta no trabalho de conclusão de curso de Alves (2012) o qual conclui que o uso de vidro ao invés de quartzo resulta em uma redução na intensidade de fluorescência e um leve deslocamento da posição dos picos de fluorescência.

Inicialmente, todas as amostras foram ambientadas a 25°C com uso de banho termostático, visando eliminar a influência da temperatura sobre os espectros de fluorescência.

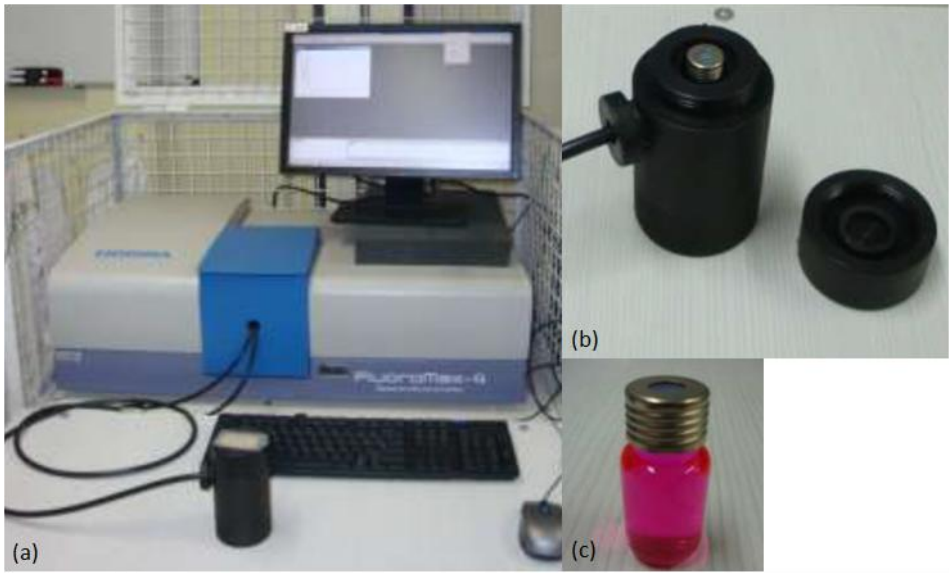

Figura 5.3: Equipamentos utilizados para a coleta dos espectros de fluorescência das amostras de diesel: (a) espectrômetro HORIBA Fluoromax®-4, com o módulo para fibra ótica; (b) câmara escura e (c) frasco utilizado para acondicionamento e medição das amostras. Fonte: Ranzan, L. (2014).

As amostras foram excitadas com comprimentos de onda variando entre 260 nm e 600 nm, com incremento de 10 nm. A emissão por sua vez foi avaliada entre os comprimentos de onda de 290 nm a 850 nm, com o mesmo incremento.

Assim, cada espectro de fluorescência conta com 1904 pares de fluorescência correspondentes à intensidade de fluorescência atribuída ao par Excitação/Emissão, distribuídos em uma matriz tridimensional onde os eixos x, y e z são respectivamente: comprimento de onda de emissão (λem), comprimento de onda de excitação (λex), e a intensidade de fluorescência. A Figura 5.4 apresenta um espectro por fluorescência típico para uma amostra de diesel HDT.

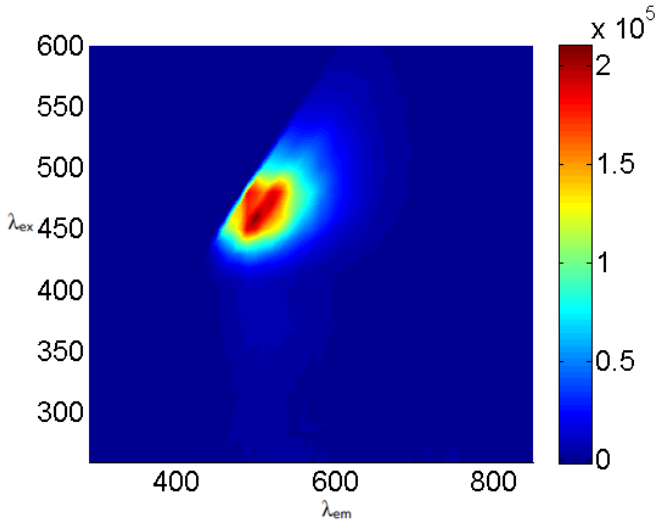

Figura 5.4: Espectro de fluorescência típico de uma amostra de diesel HDT.

5.3 Espectroscopia de Infravermelho Próximo (NIR)

A segunda técnica espectrométrica utilizada consiste na espectroscopia de infravermelho próximo, ou NIR. Os espectros das amostras de diesel foram medidos com uso do equipamento PerkinElmer Frontier FT-IR/FT-NIR, equipado com uma lâmpada halógena de quartzo que emite radições na região do ultravioleta, luz visível e, majoritariamente, do infravermelho, abrangendo o intervalo de 10000 - 4000 cm^{-1}. O equipamento consiste basicamente da fonte luminosa, de um detector e de um prisma de dispersão. As medidas foram realizadas utilizando-se um acessório denominado Near Infrared Reflectance Acessory (NIRA), também da PerkinElmer, que realiza as medições por reflectância (Figura 5.5a).

As amostras foram acondicionadas em uma placa de vidro e postas em contato com a base de uma peça hexagonal de alumínio, conforme ilustrado na Figura 5.10b. Esse item metálico, chamado de difusor, além de garantir que as amostras líquidas formem um filme com espessura uniforme, reflete o feixe de luz, o que aumenta o caminho do mesmo através da amostra. O detalhe de posicionamento da amostra no equipamento pode ser visto na Figura 5.10c.

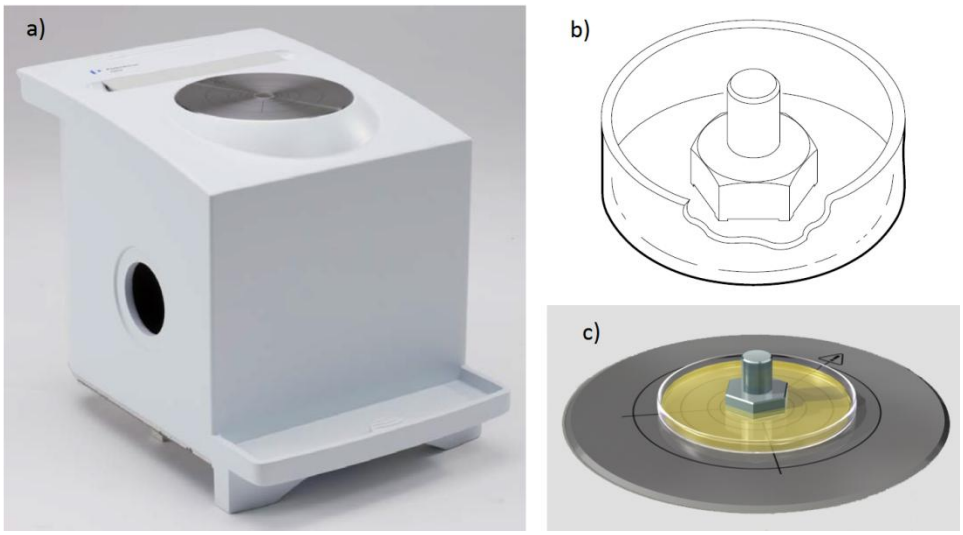

Figura 5.5: Equipamentos utilizados para a coleta dos espectros de NIR das amostras de diesel: (a) Acessório NIRA; (b) Placa de vidro e acessório metálico difusor de feixes; (c) Detalhe de posicionamento do conjunto amostral (placa + amostra + difusor) no acessório NIRA.

Os espectros foram coletados com uma diferença de 2 cm^{-1} entre cada ponto, gerando assim uma matriz de dados com 3001 colunas que correspondem à absorbância da amostra em cada comprimento de onda. Cabe salientar que os espectros obtidos foram corrigidos subtraindo-se o espectro devido à presença do difusor. A Figura 5.10 apresenta um espectro de NIR típico para uma amostra de diesel HDT.

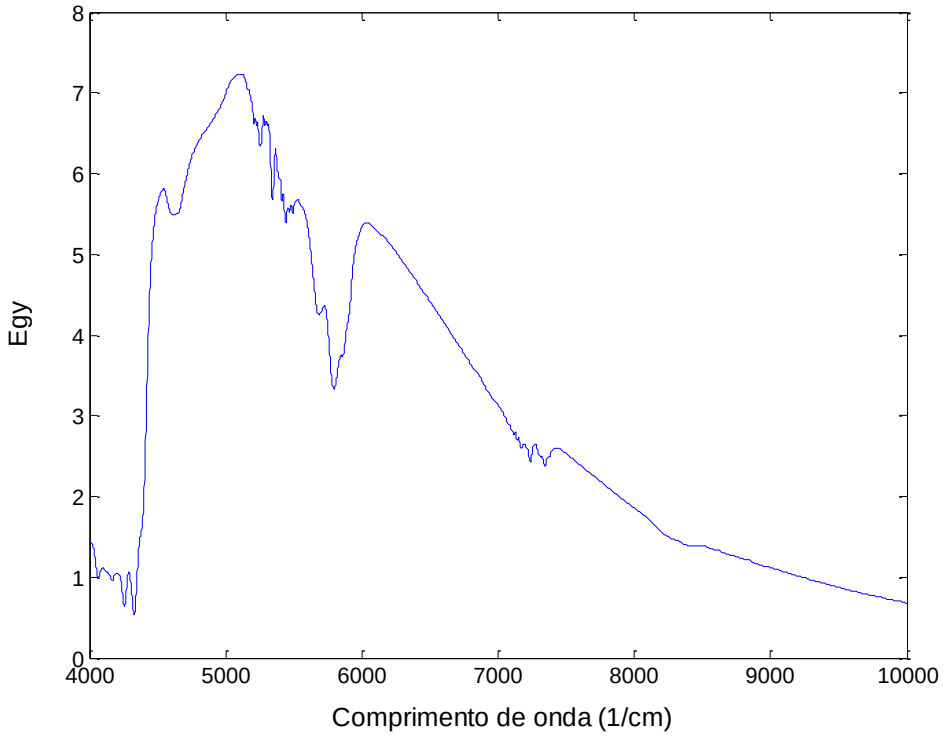

Figura 5.6: Espectro de infravermelho próximo típico de uma amostra de diesel HDT.

5.4 Avaliação do conteúdo de enxofre

A primeira etapa para avaliar o conteúdo de enxofre nas amostras de diesel consistiu na determinação do tamanho do modelo a ser construído pelos algoritmos modificados. Para isso, foi necessário analisar separadamente cada conjunto de dados de espectroscopia, realizando-se uma análise PCA para cada um. A Figura 5.7 mostra a variabilidade de dados explicada em função do número de componentes principais para cada espectroscopia. Com base nessa figura, percebe-se que a utilização de dois ou cinco componentes principais explica praticamente a mesma variabilidade dos seus respectivos dados. Assim, buscando uma equivalência de peso entre as duas espectroscopias, um menor tempo computacional e um modelo mais simplificado, optou-se por utilizar dois componentes espectrais provenientes de cada espectroscopia. Com isso, busca-se um modelo com quatro preditores no total.

Uma vez determinado o tamanho do modelo, cada uma das 28 versões do ACO (Tabela 3.1) avaliou 100 vezes o conjunto total de dados de calibração, escolhendo, em cada avaliação, um modelo formado por dois componentes da matriz de fluorescência e dois componentes da matriz de NIR. Cada avaliação utilizou cinquenta ciclos de cem formigas no algoritmo e todos os modelos envolvidos na busca foram obtidos pelo método de mínimos quadrados.

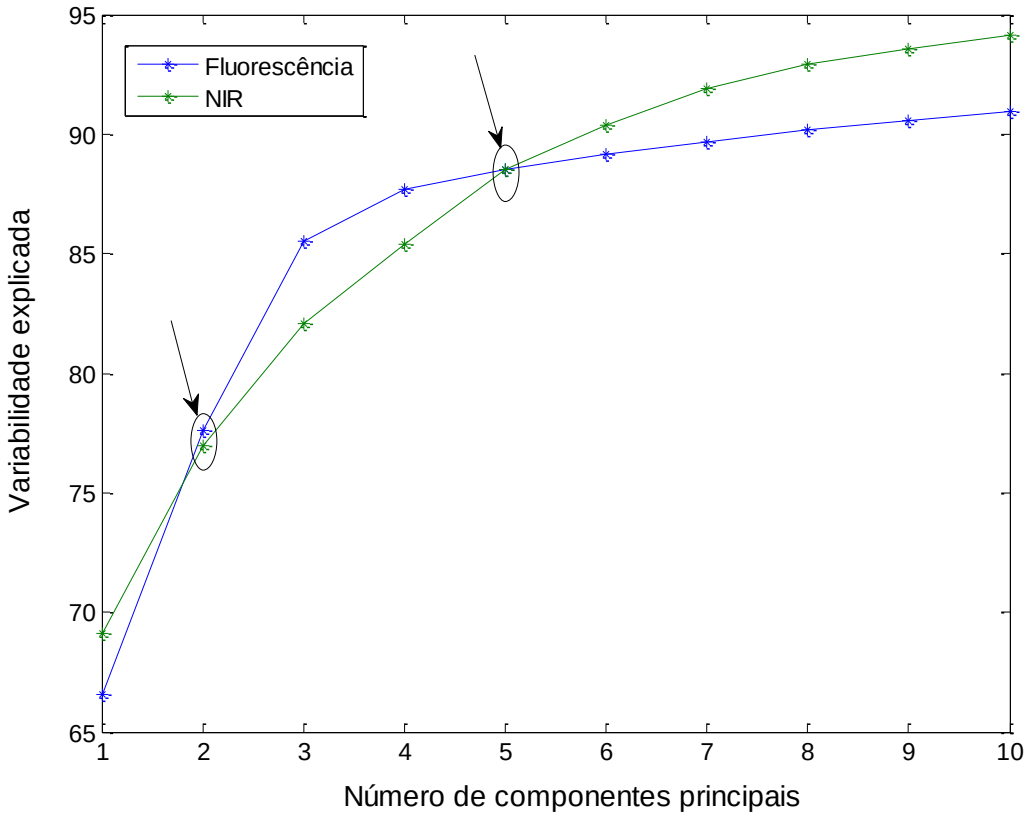

Figura 5.7: Comparação entre a variabilidade explicada de cada espectroscopia em função do número de componentes principais.

5.4.1 Discussão dos Resultados

Todos esses modelos gerados foram então aplicados ao conjunto de dados de teste, sendo computados seus valores de RMSEP. Devido à extensão dos dados, não foi realizada a busca exaustiva, sendo considerado como referência o valor mínimo encontrado para esse erro, igual a 8.4 ppm.

Inicialmente, foram comparados o número de replicações em que cada versão encontrou um valor próximo ao mínimo de referência, e o resultado pode ser visto na Figura 5.8. Percebe-se que a versão 4 encontrou os menores erros mais vezes, seguido pelas versões 21 e 28, enquanto a versão 1 (padrão) forneceu um dos piores resultados. No entanto, considerando que cada versão realizou 100 replicações, conclui-se que nenhum deles pode ser considerado satisfatório no que se refere à reprodutibilidade, pois, no melhor caso, apenas 5 replicações tiveram um erro menor que 110% o erro mínimo de referência. A meta é que esse valor seja o mais próximo possível do total de replicações realizadas.

Para avaliar a qualidade dos modelos obtidos, também é interessante determinar o ponto de corte dos erros que abrange a maioria das replicações. Neste caso, optou-se por analisar o 90º percentil dos erros de teste dos modelos gerados por cada versão, cujos valores são mostrados na Figura 5.9. Esta figura apresenta também o valor de RMSEP

encontrado pelo modelo construído por Regressão de Componentes Principais (PCR), dado pela linha contínua. Este modelo PCR é formado pelos dois primeiros componentes principais da matriz de dados de fluorescência bidimensional e os dois primeiros componentes principais da matriz de dados de NIR, resultando assim em um modelo com tamanho k=4.

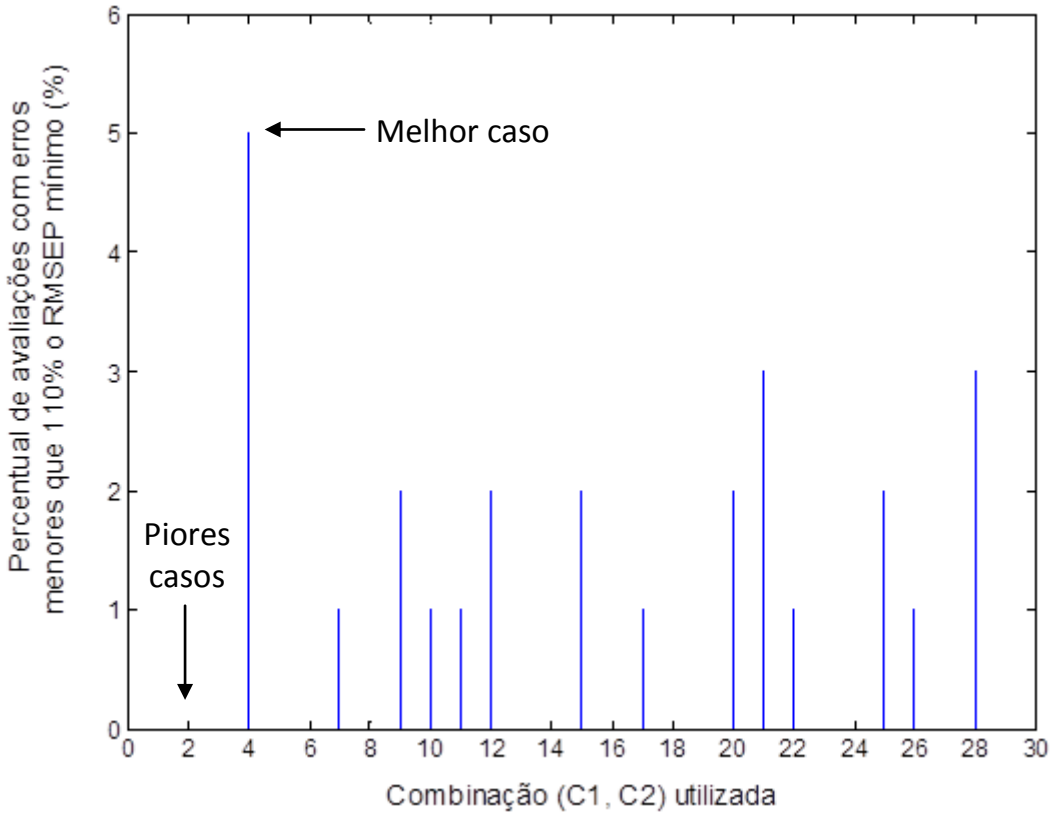

Figura 5.8: Percentual de replicações que cada par de critérios obteve um modelo com RMSEP menor ou igual a 110% o erro mínimo encontrado na fase teste.

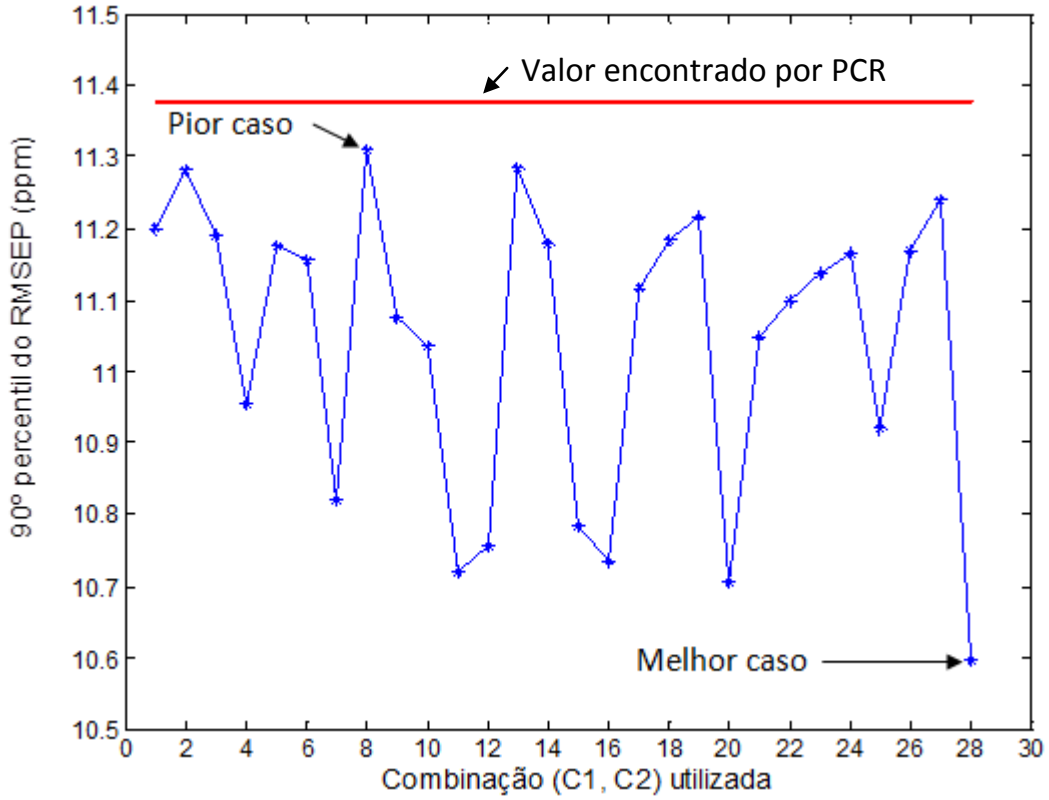

Figura 5.9: Valor do RMSEP, em ppm de enxofre, capaz de abranger 90% das replicações de cada versão (azul), e valor do RMSEP encontrado pelo modelo obtido por PCR (vermelho).

Despreende-se da Figura 5.9 que os modelos gerados utilizando-se a versão 28 forneceram os menores valores de RMSEP quando aplicados a um novo conjunto de dados. Ainda, a versão 1, considerado como o ACO padrão, apresentou um dos valores mais altos de 90º percentil, mostrando que algumas modificações no algoritmo foram, de fato, eficientes. Considerando, no entanto, a ordem de grandeza dos dados de saída, cuja média é 101 ppm, conclui-se que todas as versões fornecem erros baixos de predição. Além disso, quando comparados ao erro de predição do modelo obtido por PCR, todas as versões apresentaram um valor menor em 90% de suas replicações. Isso demonstra a melhoria de predição devido à utilização de componentes espectrais puros em detrimento de toda matriz espectral para construção de modelos.

Ainda pela Figura 5.9, percebe-se uma alternância entre os picos maiores e menores. Visto que isso se dá, sobretudo, devido às modificações no algoritmo que cada versão representa, convém analisar a influência dos dois critérios modificados. Assim, a Figura 5.11 apresenta a soma dos valores de 90º percentil dos erros encontrados pelas versõess que utilizam o mesmo C1 e a Figura 5.10 apresenta a soma dos valores de 90º percentil dos erros encontrados pelas versões que utilizam o mesmo C2.

No que diz respeito à predição do conteúdo de enxofre, os modelos gerados pela utilização do SSE na atualização das trilhas de feromônio apresentaram o pior resultado, sendo responsáveis pelos maiores erros encontrados. Por outro lado, o uso do logaritmo

de R_a^2 apresentou os menores erros de predição, mesmo sendo uma métrica de avaliação global de modelos. De fato, nesta análise, aquelas métricas consideradas melhores para otimização (LCI, Teste t, Teste F) não mostraram os melhores resultados quando aplicadas à tarefa de predição de enxofre.

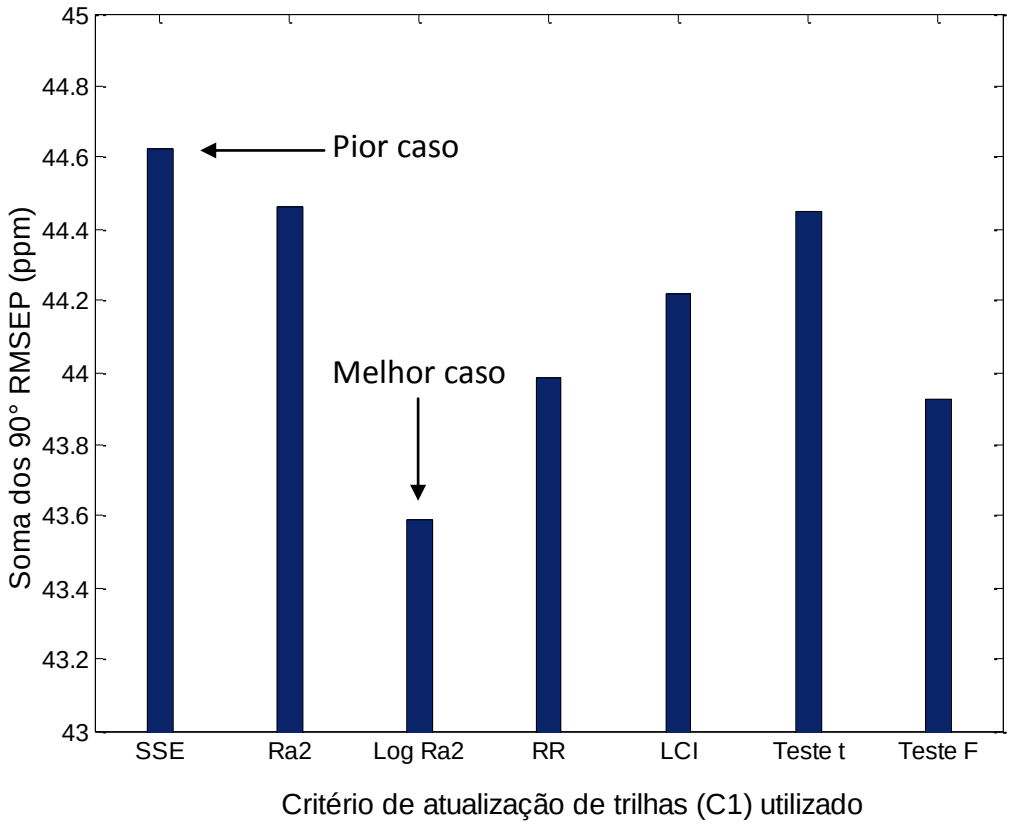

Figura 5.10: Influência de C1: soma dos valores de 90º percentil dos erros de predição encontrados pelas 4 versões que utilizam o mesmo critério de atualização de trilha.

Quanto à influêcia dos critérios de seleção de modelos, os modelos gerados pela utilização do SSE/R_a^2 apresentaram o pior resultado, sendo responsáveis pelos maiores erros encontrados, o que vai ao encontro do resultado obtido pelos estudos de caso discutidos anteriormente. Por outro lado, o uso da relação SSE/RR apresentou modelos com os menores erros de predição. Convém ressaltar também que, pelas duas últimas figuras, conclui-se que o ACO padrão (versão 1) utiliza o pior critério C1 e o segundo pior critério C2 na busca pelo melhor modelo.

Por último, é feita uma análise boxplot do conjunto de erros de predição das 28 versões, o que fornece uma informação mais aprofundada sobre a distribuição dos erros encontrados por cada versão. Neste tipo de gráfico, esta distribuição é representada por três elementos: a caixa delimitada pelo 25º percentil na base inferior e 75º percentil na base superior, abrangendo 50% dos valores centrais encontrados e segmentada por uma linha indicativa da média; os whiskers, ou bigodes, prolongamentos representados por linhas tracejadas que unem a base inferior e superior da caixa ao menor e maior elemento não considerado outlier, respectivamente; e os outliers, representados por pontos que indicam valores distantes das bases da caixa por um fator de 1.5 vezes o tamanho da mesma. No contexto deste trabalho, este último elemento é de suma importância para a avaliação do menor erro encontrado por cada versão. Para mais

informações a respeito do boxplot e sua interpretação, consultar Walpole *et al.*(2012) ou Braga (2010).

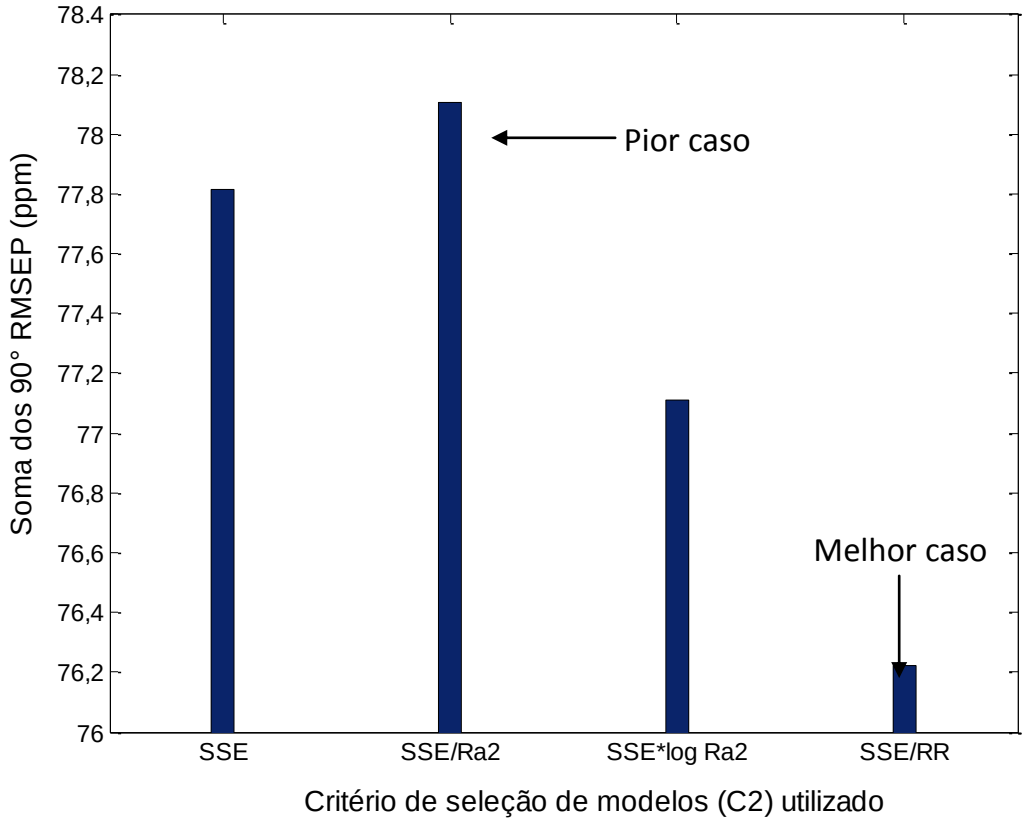

Figura 5.11: Influência de C2: soma dos valores de 90º percentil dos erros de predição encontrados pelas 7 versões que utilizam o mesmo critério de seleção de modelos.

Assim, a Figura 5.12 mostra que, por exemplo, os erros encontrados pela versão24 encontram-se bastante concentrados na região de 11 ppm, pois a caixa e os prolongamentos são bastante estreitos em torno desse valor. Por outro lado, a versão 17 apresentou uma dispersão maior dos erros de predição, representada pela caixa alongada.

Percebe-se também pela Figura 5.12 que a versão 1, que representa o ACO padrão, apresentou uma caixa relativamente "alta" e alongada quando comparado a algumas versões propostas neste trabalho, tais como o 12, 16, 20 e 28. Este último, em especial, além de apresentar o erro mínimo global também apresenta uma caixa posicionada em valores mais baixos de RMSEP, ainda que razoavelmente alongada. Além disso, é interessante evidenciar que todas as versõess encontraram um erro de predição menor que o modelo obtido por PCR (linha horizontal vermelha no valor 11.4 ppm) em praticamente todas as suas replicações.

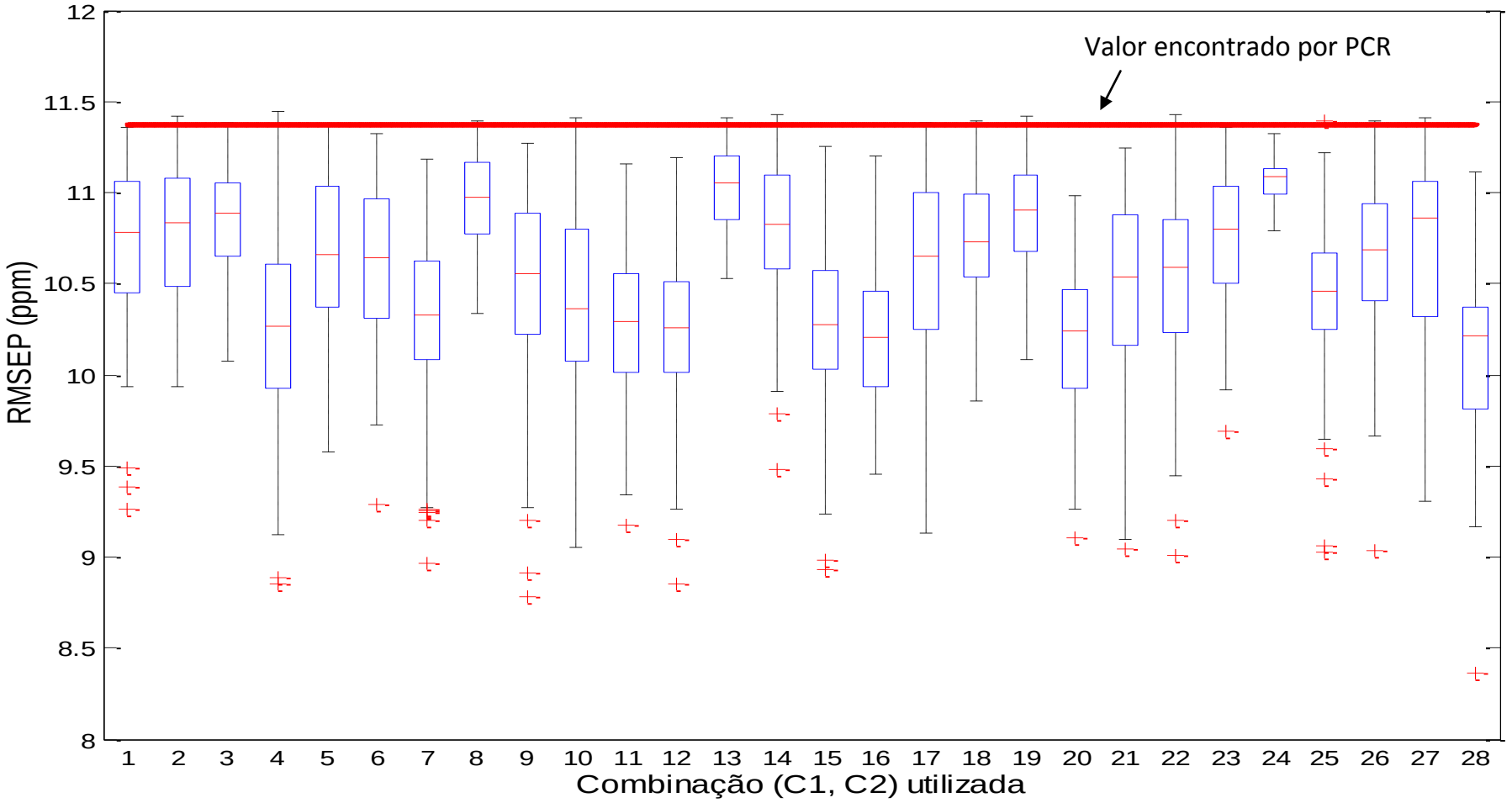

Figura 5.12: Boxplot do conjunto de valores de RMSEP encontrados pelos modelos gerados por cada combinação de critérios (C1, C2).

Além da medição da reprodutibilidade das versões, ou seja, sua capacidade de encontrar modelos com os menores erros num determinado número de replicações, é importante também definir a qualidade do modelo formado pelos componentes espectrais escolhidos com mais frequência por cada versão. Dessa forma, pode-se avaliar se determinada combinação de critérios é capaz de gerar um bom modelo com base nos preditores mais frequentemente escolhidos quando o número de replicações tende ao infinito. Isso também permite associar cada combinação a um único modelo de predição, o que facilita a análise. O resultado é apresentado na Figura 5.13.

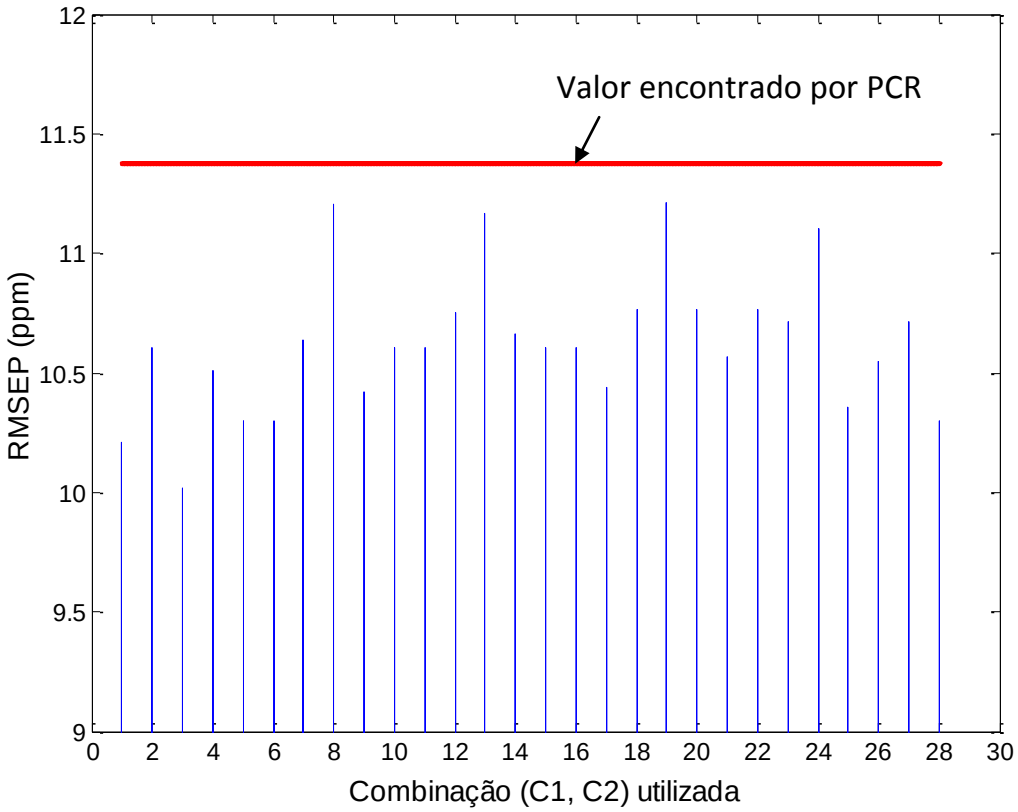

Figura 5.13: Valor do RMSEP, em ppm de enxofre, encontrado pelo modelo formado pelos componentes mais frequentes de cada combinação de critérios (azul), e valor encontrado pelo modelo obtido por PCR (vermelho).

De acordo com a Figura 5.13, os componentes encontrados com mais frequência pela versão 3 foram os que resultaram no modelo com o menor erro de predição. Assim, mesmo não tendo encontrado o erro mínimo isoladamente em nenhuma de suas replicações, conforme visto anteriormente na Figura 5.8, a tendência desta versão do ACO é escolher os melhores componentes espectrais ao longo do todas as replicações. No entanto, cabe salientar que, embora tenham fornecidos valores baixos de RMSEP, nenhuma das versões foi capaz de formar, com seus componentes mais frequentes, o modelo que gerou o erro mínimo encontrado, igual a 8.4 ppm. Mesmo assim, percebe-se

que a utilização de qualquer um deles já representa um avanço frente ao uso da regressão PCR, que utiliza toda a matriz espectral para gerar o modelo.

A última análise que deve ser feita a fim de responder todas as perguntas propostas para este trabalho é a comparação dos resultados obtidos pelo uso das duas técnicas de espectroscopia frente àqueles obtidos utilizando-se apenas a espectroscopia de fluorescência, expostos na Tabela 5.1.

Tabela 5.1: Pares de fluorescência selecionados por PSCM e valores dos erros de predição dos modelos para Diesel HDT. Fonte: Adaptado de Ranzan, L. (2014)

Número de pares	Pares Selecionados								Diesel HDT RMSEP
1	Ex450/Em460								8,5479
2	Ex410/Em470	Ex450/Em510							9,9454
3	Ex570/Em820	Ex450/Em490	Ex410/Em470						11,0114
4	Ex390/Em450	Ex300/Em350	Ex420/Em460	Ex450/Em480					8,9740
5	Ex450/Em490	Ex330/Em640	Ex410/Em460	Ex410/Em620	Ex320/Em350				8,9351
6	Ex330/Em470	Ex440/Em600	Ex430/Em470	Ex450/Em460	Ex270/Em810	Ex510/Em660			7,0648
7	Ex320/Em500	Ex410/Em520	Ex340/Em550	Ex450/Em460	Ex370/Em380	Ex280/Em550	Ex310/Em350		9,1334
8	Ex350/Em380	Ex430/Em490	Ex490/Em640	Ex450/Em460	Ex380/Em390	Ex470/Em500	Ex410/Em780	Ex410/Em470	8,4292

Os resultados mostrados na Tabela 5.1 foram retirados do trabalho de Ranzan, L. (2014), no qual foram utilizadas 51 amostras de diesel HDT, sendo que 45 são as mesmas utilizadas no presente trabalho. Além disso, os modelos foram construídos através do PSCM, que utiliza o ACO padrão (versão 1) para seleção dos pares de fluorescência e regressão multilinear para construção dos modelos, variando-se o tamanho do modelo de 1 a 8 pares de fluorescência. Não é especificado, porém, nem o número de ciclos nem o número de formigas utilizadas no algoritmo.

Pode-se observar que, utilizando apenas dois componentes do espectro de fluorescência, o menor erro encontrado foi de 9.94 ppm. Em associação com o espectro de NIR e utilizando a versão 1 foi possível obter um modelo com erro mínimo de 9,26 ppm, um valor 7% menor. No entanto, utilizando-se 4 componentes do espectro de fluorescência, formando assim um modelo do mesmo tamanho que o aqui proposto, o erro mínimo encontrado foi igual a 8,97 ppm, um valor 3% menor que o encontrado substituindo-se dois pares de fluorescência por dois componentes do espectro NIR. Isso sugere que, para essa faixa de concentração de enxofre, o espectro de fluorescência fornece mais informações que o espectro de infravermelho próximo. No entanto, o trabalho de Cramer (2009) mostra que, para altas concentrações, o espectro NIR tende a fornecer informações bastante relevantes, sendo inclusive utilizado para a construção de modelos capazes de predizer concentrações de enxofre menores que 3,7 ppb.

Capítulo 6 – Conclusões e Trabalhos Futuros

Este trabalho teve como principal objetivo a proposta de diferentes versões de seleção de variáveis com base no algoritmo ACO, modificando-se os critérios utilizados em duas de suas principais etapas: a atualização da trilha de feromônios (C1) e a seleção de modelos (C2). Foram implementados sete critérios C1 e quatro critérios C2, gerando, no total, 28 tipos de algoritmo ACO. Essas 28 versões foram comparadas a partir de dois estudos de caso, um utilizando-se espectroscopia de fluorescência para predição do conteúdo de biomassa em ambiente fermentativo e outro utilizando-se espectroscopia NIR para a predição do conteúdo de proteína em amostras de farinha.

Em relação à versão ACO padrão, sugerido por Ranzan, C. (2014) , as modificações realizadas forneceram resultados diferentes no que diz respeito à capacidade de otimização e predição de variáveis. Tanto a alteração do critério de atualização de trilhas como do critério de seleção de modelos modificaram a qualidade do algoritmo inicial, embora o segundo critério pareça ter uma influência bem menor sobre o aspecto otimizador. Em todo caso, a utilização da razão SSE/R_a^2 resulta num algoritmo que encontrou o mínimo (global ou local, dependendo do estudo de caso) com menor frequência.

Ainda, de acordo com o primeiro estudo de caso, os melhores método de otimização foram aqueles que utilizaram o tamanho do intervalo de confiança como critério para atualização de trilhas, em especial a versão 17, que utiliza o erro SSE como critério de seleção de modelos. O segundo estudo divergiu do primeiro no sentido de ter indicado a utilização do teste t como melhor critério C1. No entanto, em ambos os casos as melhores versões apontados utilizam métricas de avaliação individual dos componentes do modelo como C1. Assim, percebe-se que, de forma geral, a utilização deste tipo de métrica tende a fornecer melhores resultados de otimização que o uso de métricas de avaliação global de modelos.

Os estudos de caso apresentaram diferentes combinações ótimas de métricas, o que indica que a melhor versão depende fortemente do conjunto de dados em análise. No entanto, as melhores versões tendem a ser aquelas que utilizam métricas de avaliação

das variáveis que compõem o modelo, ou seja, o tamanho do intervalo de confiança, o teste t ou o teste F para atualização da trilha de feromônios (versões 17 a 28).

No que diz respeito à construção de modelos preditivos, todos as combinações de critério encontraram modelos com erros relativamente baixos de predição nos dois estudos. Ainda assim, conforme dito anteriormente, os erros menores estão também associados às versões que utilizam métricas de avaliação individual (versões 17 a 28).

Além disso, em todas as análises feitas, a maioria das novas versões propostas forneceu resultados melhores que a versão 1, considerada o ACO padrão. A diferença entre os resultados, no entanto, depende novamente do conjunto de dados utilizado. Além disso, essa superioridade dos algoritmos ACO modificados é mais evidente no que diz respeito à capacidade de otimização do que de construção de modelos preditivos.

As conclusões referentes à aplicabilidade das versões na caracterização de combustíveis podem ser retiradas do capítulo 5. Nele, a qualidade de cada versão na predição de enxofre em diesel foi analisada sob dois aspectos, um referente à reprodutibilidade, ou seja, a capacidade de encontrar modelos com baixos RMSEP na grande maioria das replicações, e outro referente à construção de um bom modelo a partir dos componentes selecionados com mais frequência. Em questão de reprodutibilidade, a versão 28, que utiliza o teste F como C1 e a relação SSE/RR como C2, apresentou melhores resultados, encontrando em 90% dos modelos erros de predição menores que 10,6 ppm. Em relação ao modelo formado pelos elementos espectrais mais frequentes, embora todas as versões tenham gerado erros baixos e menores que o modelo construído por PCR, a versão 3 (SSE como C1 e SSE*$|\log R_a^2|$ como C2) mostrou-se levemente superior, com um RMSEP de 10 ppm.

Finalmente, utilizando-se 4 pares de fluorescência os componentes mais frequentes selecionados pela versão 1 resultaram em um modelo com um RMSEP igual a 8,97 ppm. No entanto, ao introduzir dados de espectroscopia NIR, de forma a substituir dois componentes de fluorescência no modelo, o erro de predição sobe para 9,26 ppm. Assim, comparando os resultados encontrados pela versão 1 utilizando-se os dois conjuntos de dados espectrais frente ao uso de apenas componentes espectrais de fluorescência, conclui-se que, para um modelo de tamanho igual a 4, a inserção da espectroscopia de NIR prejudicou a predição do modelo formado.

Assim, o presente trabalho atingiu todos objetivos a que se propôs. Para trabalhos futuros, sugere-se estudar o uso de outras métricas como critérios de atualização de trilha e seleção de modelos, bem como mesclar essas diferentes métricas dentro de um mesmo algoritmo e utilizar tazas de convergência de resultados como métrica de parada dos algoritmos. Além disso, aplicar essas versões do ACO ao espectro NIR e ao espectro de fluorescência separadamente para predição de diferentes concentrações de enxofre pode evidenciar se a melhor espectroscopia depende da região de concentração. As predições de outras propriedades além do conteúdo de enxofre podem ser estudadas utilizando-se as versões propostas, a fim de verificar sua aplicabilidade ideal na tarefa de caracterização de combustíveis. Por último, pode-se estudar a possibilidade de construção de modelos não-lineares dentro do algoritmo, capazes de oferecer melhor ajuste aos dados.

Referências

Aburto, P. et al. Quantitative analysis of sulfur in diesel by enzymatic oxidation, steady-state fluorescence, and linear regression analysis. **Energy and Fuels,** v. 28, n. 1, p. 403-408, 2014.

Adams, M. J. **Chemometrics in Analytical Spectroscopy**. Wolverhampton, UK: RSC Analytical Spectroscopy Monographs, 1995.

Allegrini, F.; Olivieri, A. C. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis. **Analytica Chimica Acta,** v. 699, p. 18-25, 2011.

Alves, C. D. V. **Uma nova sistemática para análise de enxofre em diesel baseada em fluorescência.** 2012. (Trabalho de diplomação.). Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Alves, J. C. L.; Poppi, R. J. Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM). **Talanta,** v. 104, p. 155-161.

Andrade, J. M. et al. Non-destructive and clean prediction of aviation fuel characteristics through Fourier transform-Raman spectroscopy and multivariate calibration. **Analytica Chimica Acta,** v. 482, n. 1, p. 115-128, 2003.

Agência Nacional Do Petróleo, Gás Natural E Biocombustíveis (ANP). **Resolução ANP Nº 50**. Brasil: DOU 24.12.2013.

Araki, T.; Ikeda, K.; Akaho, S. An efficient sampling algorithm with adaptations for Bayesian variable selection. **Neural Networks,** v. 61, p. 22-31, 2015.

ASTM D6258-09(2014), **Standard Test Method for Determination of Solvent Red 164 Dye Concentration in Diesel Fuels**, ASTM International, West Conshohocken, PA, 2014.

ASTM D2622-10, **Standard Test Method for Sulfur in Petroleum Products by Wavelength Dispersive X-ray Fluorescence Spectrometry**, ASTM International, West Conshohocken, PA, 2010.

ASTM D4294-10, **Standard Test Method for Sulfur in Petroleum and Petroleum Products by Energy Dispersive X-ray Fluorescence Spectrometry**, ASTM International, West Conshohocken, PA, 2010.

ASTM D5453-12, **Standard Test Method for Determination of Total Sulfur in Light Hydrocarbons, Spark Ignition Engine Fuel, Diesel Engine Fuel, and Engine Oil by Ultraviolet Fluorescence**, ASTM International, West Conshohocken, PA, 2012.

ASTM D7039-13, **Standard Test Method for Sulfur in Gasoline, Diesel Fuel, Jet Fuel, Kerosine, Biodiesel, Biodiesel Blends, and Gasoline-Ethanol Blends by Monochromatic Wavelength Dispersive X-ray Fluorescence Spectrometry**, ASTM International, West Conshohocken, PA, 2013.

ASTM D7212-13, **Standard Test Method for Low Sulfur in Automotive Fuels by Energy-Dispersive X-ray Fluorescence Spectrometry Using a Low-Background Proportional Counter**, ASTM International, West Conshohocken, PA, 2013.

ASTM D7220-12, **Standard Test Method for Sulfur in Automotive, Heating, and Jet Fuels by Monochromatic Energy Dispersive X-ray Fluorescence Spectrometry**, ASTM International, West Conshohocken, PA, 2012.

Balabin, R. M.; Lomakina, E. I.; Safieva, R. Z. Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. **Fuel,** v. 90, n. 5, p. 2007-2015, 2011.

Balabin, R. M.; Safieva, R. Z. Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data. **Analytica Chimica Acta,** v. 689, n. 2, p. 190-197, 2011.

Baptista, P. et al. Multivariate near infrared spectroscopy models for predicting the iodine value, CFPP, kinematic viscosity at 40°C and density at 15°C of biodiesel. **Talanta,** v. 77, n. 1, p. 144-151, 2008.

Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. **Chemometrics: A pratical guide**. New York: Wiley & Sons, 1998.

BRAGA, L. P. V. **Compreendendo Probabilidade e Estatística**. Rio de Janeiro: E-papers, 2010.

Breitkreitz, M. C. et al. Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. **Analyst,** v. 128, n. 9, p. 1204-1207, 2003.

Brereton, G. R. **Chemometrics: Data Analysis for the Laboratory and Chemical Plant.** Bristol: John Wiley & Sons Ltd, 2003.

Brown, S. D.; Tauler, R.; Walczak, B. Comprehensive chemometrics: Chemical and biochemical data analysis. **Analytical and Bioanalytical Chemistry**, p. 1-2, 2009.

Burns, D. A.; Ciurczak, E. W. **Handbook of Near-Infrared Analysis**. 2nd. New York: Marcel Dekker, 2001.

Canha, N. et al. Multivariate near infrared spectroscopy models for predicting the oxidative stability of biodiesel: Effect of antioxidants addition. **Fuel,** v. 97, p. 352-357, 2012.

Cerny, V. **Journal of Optimization Theory and Applications**, 1985.

Cramer, J. A. et al. Ultra-low sulfur diesel classification with near-infrared spectroscopy and partial least squares. **Energy and Fuels,** v. 23, n. 2, p. 1132-1133, 2009.

de Lira, L. d. F. B. et al. Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel. **Microchemical Journal,** v. 96, n. 1, p. 126-131, 2010.

de Lira, L. d. F. B. et al. Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration. **Fuel,** v. 89, n. 2, p. 405-409, 2010.

Dorigo, M.; Gambardella, L. M. Ant Colonies for the Tavelling Salesman Problem. **Biosystems,** v. 43, p. 73 - 81, 1997.

Felizardo, P. et al. Multivariate near infrared spectroscopy models for predicting methanol and water content in biodiesel. **Analytica Chimica Acta,** v. 595, p. 107-113, 2007.

Ferd, W. Review of present trends in luminescence research. **Journal of Luminescence,** v. 24-25, Part 2, p. 929-936, 1981.

Ferrão, M. F. et al. Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. **Fuel,** v. 90, n. 2, p. 701-706, 2011.

Geladi, P. Chemometrics in spectroscopy. Part 1. Classical chemometrics. **Spectrochimica Acta Part B: Atomic Spectroscopy,** v. 58, n. 5, p. 767-782, 2003.

Geladi, P. et al. Chemometrics in spectroscopy: Part 2. Examples. **Spectrochimica Acta Part B: Atomic Spectroscopy,** v. 59, n. 9, p. 1347-1357, 2004.

Ghasemi-Varnamkhasti, M. et al. Screening analysis of beer ageing using near infrared spectroscopy and the Successive Projections Algorithm for variable selection. **Talanta,** v. 89, p. 286-291, 2012.

Goicoechea, H. C.; Olivieri, A. C. A New Family of Genetic Algorithms for Wavelengh Intervals Selection in Multivariate AnalyticalSpectroscopy. **Journal of Chemometrics.** v. 17**,** p. 338-345, 2003.

Goldberg, D. E. **Genetic Algorithms in Search, Optimization and Machine Learning**. Boston: Kluwer Academic Publishers, 1989.

Goodarzi, M.; dos Santos Coelho, L. Firefly as a novel swarm intelligence variable selection method in spectroscopy. **Analytica Chimica Acta,** v. 852, p. 20-27, 2014.

Goss, S. et al. Self-organized Shortcuts in the Argentine Ant. **Naturwissenschaften,** v. 76, p. 579 - 581, 1989.

Hantelmann, K. et al. Two-dimensional fluorescence spectroscopy: a novel approach for controlling fed-batch cultivations. **Journal of Biotechnology,** v. 121, p. 410-417, 2006.

Hapfelmeier, A.; Ulm, K. Variable selection by Random Forests using data with missing values. **Computational Statistics & Data Analysis,** v. 80, p. 129-139, 2014.

Hemmateenejad, B. et al. Building optimal regression tree by ant colony system–genetic algorithm: Application to modeling of melting points. **Analytica Chimica Acta,** v. 704, n. 1–2, p. 57-62, 2011.

Hitzmann, B. et al. **Chemometric models for the on-line estimation of bioprocess variables from 2-D fluorescencespectra**. 7th International Conference on Computer Applications in Biotechnology. Osaka, 1998.

Kara, S. et al. Fluorescence spectroscopy as a novel method for on-line analysis of biocatalytic C–C bond formations. **Journal of Molecular Catalysis B: Enzymatic,** v. 66, p. 124-129, 2010.

Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by simulated annealing. **Science**, 1983.

Knüttel, T. et al. On-line monitoring of a quasi-enantiomeric reaction with two coumarin substrates via 2D-fluorescence spectroscopy. **Enzyme and Microbial Technology,** v. 29, n. 2–3, p. 150-159, 2001.

Kumar, N. et al. Chemometrics tools used in analytical chemistry: An overview. **Talanta**, 2014.

Lakowicz, J. R. **Principles of fluorescence spectroscopy**. 3rd. New York: Springer, 2006.

Leardi, R.; Seasholtz, M. B.; Pell, R. J. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. **Analytica Chimica Acta,** v. 461, n. 2, p. 189-200, 2002.

Li, J.; Dai, L. A hard modeling approach to determine methanol concentration in methanol gasoline by Raman spectroscopy. **Sensors and Actuators B: Chemical,** v. 173, p. 385-390, 2012.

Li, S.; Dai, L.-k. Classification of gasoline brand and origin by Raman spectroscopy and a novel R-weighted LSSVM algorithm. **Fuel,** v. 96, p. 146-152, 2012.

Lindemann, C. et al. 2-Dimensional fluorescence spectroscopy for on-line bioprocess monitoring. **Sensors and Actuators B: Chemical,** v. 51, n. 1-3, p. 273-277, 1998.

Marose, S.; Lindemann, C.; Scheper, T. Two-dimensional fluorescence spectroscopy: A new tool for on-line bioprocess monitoring. **Biotechnology Progress,** v. 14, n. 1, p. 63-74, 1998.

Medeiros, A. R. B. **Uso de ATR/FTIR e FTNIR associado a técnicas quimiométricas para quantificação de aditivos em gasolina automotiva.** 2009. Instituto de Química Universidade de Brasília, DF, Brasil.

Mello, P. A.; Pinto, J. C. C. S. **Introdução à Modelagem Matemática e Dinâmica Não-Linear de Processos Químicos**. Rio de Janeiro: *Escola Piloto Virtual Giuliano Massaran*, 2008.

Mendes, L. S. et al. Determination of ethanol in fuel ethanol and beverages by Fourier transform (FT)-near infrared and FT-Raman spectrometries. **Analytica Chimica Acta,** v. 493, n. 2, p. 219-231, 2003.

Mulchandani, A.; Bassi, A. S. Principles and applications of biosensors for bioprocess monitoring and control. **Crit. Rev. Biotechnol,** v. 1, p. 105-124, 1995.

Mullen, R. J. et al. A review of ant algorithms. **Expert Systems with Applications,** v. 36, n. 6, p. 9608-9617, 2009.

Oliveira, F. R. P. et al. Chemometric modelling for process analyzers using just a single calibration sample. **Chemometrics and Intelligent Laboratory Systems,** v. 94, n. 2, p. 118-122, 2008.

Omary, M. A.; Patterson, H. H. Luminescence, Theory. In: Editor-in-Chief: John, L. (Ed.). **Encyclopedia of Spectroscopy and Spectrometry (Second Edition)**. Oxford: Academic Press, 1999. p.1372-1391.

Paiva, D. L.; Lampman, G. M.; Kriz, G. S. **Introduction to spectroscopy**. 3rd. Thomson Learning, 2001.

Pantoja, P. A. et al. Prediction of crude oil properties and chemical composition by means of steady-state and time-resolved fluorescence. **Energy and Fuels,** v. 25, n. 8, p. 3598-3604, 2011.

Parkash, S. **Petroleum Fuels Manufacturing Handbook.** McGraw Hill, 2010.

Pasquini, C. **Espectroscopia no Infravermelho Proximo (NIR)**. Salvador: UFBA 2002.

Pattison, R. N. et al. Measurement and control of dissolved carbon dioxide in mammalian cell culture processes using an in situ fiber optic chemical sensor. **Biotechnology progress,** v. 16, p. 769-774, 2000.

Pilar Dorado, M. et al. Visible and NIR Spectroscopy to assess biodiesel quality: Determination of alcohol and glycerol traces. **Fuel,** v. 90, n. 6, p. 2321-2325, 2011.

Ranzan, C. **Desenvolvimento De Modelos Quimiométricos Utilizando O Algoritmo De Otimização Colônia De Formigas**. 2014. (PhD). Chemical Engineering, Federal University of Rio Grande do Sul, Porto Alegre.

Ranzan, C. et al. Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization. **Chemometrics and Intelligent Laboratory Systems,** v. 132, p. 133-140, 2014.

Ranzan, L. **Estudo da Viabilidade do Uso de Espectroscopia Fluorescente 2D para Quantificar Reor de Enxofre em Óleo Diesel**. 2014. (MSc). Chemical Engineering, Federal University of Rio Grande do Sul, LUME - UFRGS.

Ranzan, L. **Estudo Da Viabilidade Do Uso De Espectroscopia Por Fluorescência 2d Para Quantificar Teor De Enxofre Em Óleo Diesel**. 2014. (Msc). Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Riveros, L. et al. Determination of asphaltene and resin content in Venezuelan crude oils by using fluorescence spectroscopy and partial least squares regression. **Energy and Fuels,** v. 20, n. 1, p. 227-230, 2006.

Santos Jr, V. O. et al. A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. **Analytica Chimica Acta,** v. 547, n. 2, p. 188-196, 2005.

Scheper, T. et al. Bioanalytics: detailed insight into bioprocesses. **Analytica Chimica Acta,** v. 400, p. 121-134, 1999.

Scherer , M. D. et al. Determination of the Biodiesel Content in Diesel/BiodieselBlends: A Method Based on Fluorescence Spectroscopy. **Journal of Fluorescence**. v. 21**,** p. 1027-1031, 2011.

Silva, M. A. A. et al. A new spectrophotometric method for determination of biodiesel content in biodiesel/diesel blends. **Fuel,** v. 143, p. 16-20, 2015.

Silveira, R. **Novas Metodologias Para Compressão de Dados de Processos e Sistemáticas Para Ajuste do Sistema PI**. 2012. (Msc.). Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Skibsted, E. et al. On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration. **Journal of Biotechnology,** v. 88, n. 1, p. 47-57, 2001.

Skoog, D. A.; Holler, F. J.; Crouch, S. R. **Principles of instrumental analysis**. 6th. Belmont, CA: Thomson Brooks/Cole, 2007.

Socha, K.; Dorigo, M. Ant colony optimization for continuous domains. **European Journal of Operational Research,** v. 185, n. 3, p. 1155-1173, 2008.

Solle, D. et al. Chemometric Modelling based on 2D-Fluorescence Spectra without a Calibration Measurement. **Bioinformatics,** v. 19, p. 173-177, 2003.

Sotomayor, M. D. P. T. et al. Aplicação e avanços da espectroscopia de luminescência em análises farmacêuticas. **Química Nova,** v. 31, n. 7, p. 1755-1774, 2008.

Speight, J. G. **Handbook of Petroleum Product Analysis** New Jersey: John Wiley & Sons, 2002.

Speight, J. G. **Synthetic fuels handbook**. McGraw Hill, 2008.

Spiegelman, C. H. et al. Theoretical Justification of Wavelengh Selection in PLS Calibration: Development of a New Algorithm. **Analytica Chimica Acta** v.70**,** p. 35-44, 1998.

Stärk, E. et al. In-Situ-fluorescence-probes: A useful tool for non-invasive bioprocess monitoring. **Advances in biochemical engineering Biotechnology,** v. 74, p. 21 - 38, 2002.

Sun, D.-W. **Infrared Spectroscopy for Food Quality Analysis and Control**. Elsevier, 2009.

Sykes, A. O. **An Introduction to Regression Analysis**. Law and Economics Working Paper. v. 20, p**.** 1-34, p. 1993.

Walpole, R. E. et al. **Probability & Statistics for Engineers & Scientists.** 9th. Boston: Pearson, 2012.

Wang, X. et al. Analysis on fluorescence of dual excitable Eu(TTA)3DPBT in toluene solution and PMMA. **Journal of Luminescence,** v. 131, n. 8, p. 1719-1723, 2011.

Wehrens, R.; SpringerLink. **Chemometrics with R Multivariate Data Analysis in the Natural Sciences and Life Sciences**. Use R. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2011.

Wilcox, R. **Introduction to Robust Estimation and Hypothesis Testing.** 3rd. Elsevier, 2012.

Wong, W. C. et al. Miniature pH optical fiber sensor based on waist-enlarged bitaper and mode excitation. **Sensors and Actuators B: Chemical,** v. 191, p. 579-585, 2014.

Yamuna, R. K.; Ramachandra, R. V. S. Control of fermenters – a review. **Bioprocess Engineering,** v. 21, p. 77-88, 1999.

Zhang, W.; Wang, X.; Chen, L. Improved leaps and bounds variable selection algorithm based on principal component analysis. **Chemometrics and Intelligent Laboratory Systems,** v. 139, p. 76-83, 2014.

Zhang, W. et al. Predicting the dynamic and kinematic viscosities of biodiesel–diesel blends using mid- and near-infrared spectroscopy. **Applied Energy,** v. 98, p. 122-127, 2012.

Župerl, Š. et al. Experimental determination and prediction of bilitranslocase transport activity. **Analytica Chimica Acta,** v. 705, p. 322-333, 2011.

Anexos

Anexo I-Especificações brasileiras para óleo diesel rodoviário. Fonte: Resolução ANP 50 (2011).

CARACTERÍSTICA (1)	UNIDADE	LIMITE				MÉTODO	
		TIPO A e B				ABNT NBR	ASTM/EN
		S10	S50 (2)	S500	S1800 (3)		
Aspecto	-	Límpido e isento de impurezas				14954 (4)	D4176
							-4
Cor	-	-5		-6	-7	-	-
Cor ASTM, máx.	-	3,0 (8)				14483	D1500
							D6045
Teor de biodiesel (9)	% volume	-10				15568	EN 14078
Enxofre total, máx.	mg/kg	10	50	-	-	-	D2622
		-24				-	D5453
						-	D7220
						-	D7212 (11)
						-	
		-		500	1800	-	D2622
						14533	D4294
						-	D5453
							D7220
Massa específica a 20°C	kg/m³	820 a 850 (12)		820 a 865	820 a 880	7148	D1298
						14065	D4052
Ponto de fulgor, mín.	°C	38				7974	D56
						14598	D93
						-	D3828
Viscosidade a 40°C	mm²/s	2,0 a 4,5	2,0 a 5,0			10441	D445
Destilação							
10% vol., recuperados,	°C	180 (mín.)	Anotar			9619	D86
50% vol., recuperados		245,0 a 295,0	245,0 a 310,0				
85% vol., recuperados, máx.		-		360	370		
90% vol., recuperados		-	360,0 (máx.)	Anotar			
95% vol., recuperados, máx		370	-				
Ponto de entupimento de filtro a frio, máx.	°C	-13				14747	D6371
Número de cetano, mín. ou	-	48	46	42 (14)		-	D613
Número de cetano derivado						-	D6890
(NCD), mín.						-	D7170
Resíduo de carbono Ramsbottom no resíduo dos 10% finais da destilação, máx.	% massa	0,25				14318	D524
Cinzas, máx.	% massa	0,01				9842	D482
Corrosividade ao cobre, 3h a 50°C, máx.	-	1				14359	D130
Teor de água (15), máx.	mg/kg	200	Anotar	500	-		D6304
							EN ISO
							12937
Contaminação total (15)	mg/kg	24 (máx.)	Anotar	-	-	-	EN 12662
Água e sedimentos, máx. (16)	% volume	0,05				-	D2709
Hidrocarbonetos policíclicos aromáticos, máx. (17)	% massa	11	Anotar	-	-	-	D5186
							-18
							D6591 (18)
							EN 12916
							-18
Estabilidade à oxidação (17), máx.	mg/100mL	2,5	Anotar	-	-	-	D2274 (19)
							D5304
Índice de neutralização	mg KOH/g	Anotar		-	-	14248	D974
Lubricidade, máx.	µm	-20			-	-	D 6079
							ISO
							12156
Condutividade elétrica, mín.	pS/m	25		25	Anotar	-	D2624
		-21		(21)(22)	-23	-	D4308

Apêndices

Apêndice I- Implementação da função “acow” em Matlab.

```
function [combmin,erromin,parmin,iteracoes,trail] =
acow(Vdata,Sdata,nw,niter,nform,feromon,criterio,funcao,ro,model,Q,tau0,s
elecao)

n = size(Sdata,2);

if nargin <= 1 || isempty(Vdata) || isempty(Sdata)
   error('Verificar as entradas')
   return
end

if nargin == 2
   nw = 3;
   niter = 50;
   nform = 100;
   feromon = 0;
   criterio =[0,0];
   funcao = 'linajust';
   ro = 0.5;
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 3
   niter = 50;
   nform = 100;
   feromon = 0;
   criterio =[0,0];
   funcao = 'linajust';
   ro = 0.5;
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 4
   nform = 100;
   feromon = 0;
   criterio =[0,0];
   funcao = 'linajust';
   ro = 0.5;
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 5
   feromon = 0;
   criterio =[0,0];
```

```
   funcao = 'linajust';
   ro = 0.5;
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 6
   criterio =[0,0];
   funcao = 'linajust';
   ro = 0.5;
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 7
   funcao = 'linajust';
   ro = 0.5;
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 8
   ro = 0.5;
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 9
   model = 'Modelo_lin';
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 10
   Q = 100;
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 11
   tau0 = 1e-6;
   selecao = 1;
end

if nargin == 12
   selecao = 1;
end
%=====================
if isempty(funcao)
   funcao = 'linajust';
end
if isempty(niter)
    niter = 50;
end
```

```
if isempty(nform)
    nform = 100;
end
if isempty(ro)
    ro = 0.5;
end
if isempty(model)
    model = 'Modelo_lin';
end
if isempty(Q)
    Q = 100;
end
if isempty(tau0)
    tau0 = 1e-6;
end
if isempty(selecao)
    selecao = 1;
end
if isempty(nw)
    nw = 3;
end
if isempty(feromon)
    feromon = 0;
end
if isempty(criterio)
    criterio = [0,0];
end

% implementação do algoritmo

tau(1,1:n) = tau0 ; % trilha inicial
trail = [];
trail = tau;
%Inicializando variaveis de resposta - sem otimizacao
[combmin,erromin,R2amin,logR2amin,RRmin,tmin,intervalomin,parmin,stepsmin
] = feval(funcao,Vdata,Sdata,tau,n,model,selecao,nw); % acha o erro a%
partir de uma rota aleatoria

combmin_evolucao=[];
for iter=1:niter
   for form = 1:nform

[comb(form,:),erro(form),R2a(form),logR2a(form),RR(form),t(form,:),interv
alo(form,:),par(:,form),steps(form,:)] =
feval(funcao,Vdata,Sdata,tau,n,model,selecao,nw);
   end
      tau = tau*ro;
   for form = 1:nform
       for i=1:nw

           switch criterio(1)
              case 0; tau(comb(form,i)) = tau(comb(form,i)) +
Q/(erro(form));
              case 1; tau(comb(form,i)) = tau(comb(form,i)) +
Q*R2a(form);
              case 2; tau(comb(form,i)) = tau(comb(form,i)) +
Q/logR2a(form);
              case 3; tau(comb(form,i)) = tau(comb(form,i)) +
Q*abs(RR(form));
```

```
              case 4; tau(comb(form,i)) = tau(comb(form,i)) +
Q/(erro(form)*intervalo(form,i));
              case 5;  tau(comb(form,i)) = tau(comb(form,i)) +
Q*(abs(t(form,i)))/(erro(form));
              case 6; tau(comb(form,i)) = tau(comb(form,i)) +
Q*(steps(form,i));
            end
       end

       if feromon == 1
          trail = [trail; tau];
      else
          trail = tau;
       end

   end

   switch criterio(2)
       case 0;
           if min(erro) < erromin
              [erromin,ind] = min(erro);
               combmin = comb(ind,:);
               parmin  = par(:,ind);
               tmin = t(ind,:);
               intervalomin=intervalo(ind,:);
               iteracoes=iter;
               logR2amin=logR2a(ind);
               case0=[R2a(ind) erromin iter ind];
           end

       case 1
        varsel=erro./R2a;
         if min(varsel)<(erromin./R2amin);
           [varsel,ind]=min(varsel);
           erromin = erro(ind);
           combmin = comb(ind,:);
              parmin  = par(:,ind);
               tmin = t(ind,:);
               intervalomin=intervalo(ind,:);
                iteracoes=iter;
                logR2amin=logR2a(ind);
               case1=[R2a(ind) erromin iter ind];

            end;

       case 2
           varsel=erro.*logR2a;
           if min(varsel) < (erromin*logR2amin)
               [vsmin,ind] = min(varsel);
               erromin=erro(ind);
               combmin = comb(ind,:);
               parmin  = par(:,ind);
               logR2amin=logR2a(ind);
               iteracoes=iter;
               case2 = [R2a(ind), erro(ind), iter, ind];
           end;

           case 3
            varsel=erro./abs(RR);
            if min(varsel) < (erromin./abs(RRmin))
               [vsmin,ind] = min(varsel);
               RRmin=RR(ind);
               erromin=erro(ind);
```

```
                combmin = comb(ind,:);
                parmin  = par(:,ind);
                logR2amin=logR2a(ind);
                iteracoes=iter;
                case3 = [R2a(ind), erro(ind), iter, ind];
            end;
        end;

end;
```

Apêndice II- Implementação da função “linajust” em Matlab.

```
function [listav,erro,R2a,logR2a,RR,t,intervalo,par,steps] =
linajust(Vdata,Sdata,tau,n,model,sorteio,nw)

listanv = 1:n;
listav = [];

lista(1) = ceil(n*rand(1));
listav = lista(1);
deltaf=-1;
R=corrcoef(Sdata);
retirar=[];

for i=1:size(lista,2)
    listanv = remc(listanv,lista(i));
end
erro = 0;


for i = size(lista,2)+1:nw
   for j = 1:length(listanv)
      p(j) = tau(listanv(j));
      if isnan(p(j));
          p(j)=0;
      end
   end
   switch sorteio
   case 1
      p = p./sum(p);
      fatia(1) = 0;
      for k = 2:n-i+2
         fatia(k) = fatia(k-1) + p(k-1);
      end

      pcs = find(fatia>rand(1));
      iterations=0;
     while numel(pcs)==0 && iterations<5
        pcs = find(fatia>rand(1));
        iterations=iterations+1;
         end

      pc = listanv(pcs(1) - 1);
    case 0
      npcs = find(p == max(p));
      pc = listanv(npcs(1));
   otherwise
      error('Informe corretamente a seleçao!')
      break
   end
   listav(i) = pc;
   listanv = remc(listanv, find(listanv == listav(i)));

   clear p fatia

end

switch model
    case 'Modelo_lin'
```

```
[erro,R2a,logR2a,RR,t,intervalo,par,ajt,R2,RMSEP,X,SST]=Modelo_lin(Vdata,
Sdata,listav);
        X_step=Sdata(:,listav);

        p=nw; j=nw-1;n=size(Vdata,1);alfa=0.05;
        for i=1:nw
            listav_sub= remc(listav,i);
            [erro_sub]=Modelo_lin(Vdata,Sdata,listav_sub);
           F(i)=((erro_sub-erro)/(p-j))/(erro/(n-p-1));
           if F(i)> finv(1-alfa,p-j,n-p-1);
           decisao(i)=1;
           else
           decisao(i)=0;
           end;
        s(i)=F(i)*decisao(i);
        end;

        ss=s./erro;
        steps=s./erro;
    otherwise
        error('Informe corretamente o modelo a ser usado!')
        return
end
```

Apêndice III- Implementação da função "Modelo_lin" em Matlab.

```
function
[erro,R2a,logR2a,RR,t,intervalo,par,ajt,R2,RMSEP,X,SST]=Modelo_lin(Vdata,
Sdata,wl,par)

% Este modelo é estruturado de forma linear e correlaciona as variaveis
% independentemente.

%     ENTRADAS
%      Vdata: Vetores coluna em variaveis de estado
%      Sdata: Matrix de dados espectrais
%      wl: Vetor com a possicao das variaveis independentes. Repare que
%          esta possicao é relativa a matriz apenas de variaveis
independentes!
%          Nesta matriz nao entra o padrao avaliado!!
%     SAIDAS
%      erro: Somatorio dos erros entre o padrao e a predicao
%      par:  Parametros ajustados no modelo
%      ajt:  Vetor de predicao
%      R2:   Indice de correlacao ao quadrado entre padrao e predicao
%      RMSEP: Raiz quadrada das diferencas medias


ybar=mean(Vdata); % Media da variavel de interesse
SST=(Vdata-ybar)'*(Vdata-ybar); % Soma dos quadrados totais
n=size(Vdata,1); %Número de amostras
k=length(wl); %Tamanho do modelo
nw=length(wl); %Tamanho do modelo
bj0=0; %Hipótese nula (coeficiente não significativo)

% Normalizando Variaveis de estado:
norm=[];
for i=1:size(Vdata,2)
    norm = [norm Vdata(:,i)./max(Vdata(:,i))];
end
X=[Sdata(:,wl) ones(size(Sdata,1),1)]; %Matriz de entradas
c=inv(X'*X);


if size(Vdata,2)>1
    % MIMO
    if nargin == 3
        erro=0;
        for i=1:size(Vdata,2)
            par(:,i) = inv(X'*X)*X'*norm(:,i);
            ajt(:,i) = X*par(:,i);
            E        = ajt(:,i) - norm(:,i);
            erro     = erro + E'*E;
            RMSEP(i) = sqrt(E'*E/size(Vdata,1));
            RHO      = corrcoef(norm(:,i),ajt(:,i));
            R2(i)    = RHO(1,2)^2;
        end
    end
    if nargin == 4
        if length(wl)+1 ~= size(par,1) || size(Vdata,2) ~= size(par,2)
            error('Entre vetores de parametros e de Variaveis corretos!')
            return
        end
        for i=1:size(Vdata,2)
            ajt(:,i) = X*par(:,i);
```

```
            E        = ajt(:,i) - norm(:,i);
            erro(i)  = E'*E;
            RMSEP(i) = sqrt(erro(i)/size(Vdata,1));
            % R²
            SQT      = norm(:,i)-mean(norm(:,i));% Usa dados da medicao
            SQT      = SQT'*SQT; % Usa dados da medicao
            R2(i)    = 1-(erro(i)/SQT);
        end
    end

else
    %MISO
    if nargin == 3
        par   = inv(X'*X)*X'*Vdata;
        ajt   = X*par;
        E     = ajt - Vdata;

        % SSE
        erro  = E'*E;

        RMSEP = sqrt(erro/size(Vdata,1));
        RHO   = corrcoef(Vdata,ajt);
        R2    = RHO(1,2)^2;

    end

    if nargin == 4
        if length(wl)+1 ~= length(par)
            error('Entre vetores de parametros e de Variaveis corretos!')
            return
        end
        ajt = X*par';
        E   = ajt - Vdata;
        erro= E'*E;
        R2a=1-(erro/(n-k-1))/(SST/(n-1));
        logR2a=abs(log(R2a));
        RMSEP = sqrt(erro/size(Vdata,1));
        % R²
        RHO   = corrcoef(Vdata,ajt);
        R2    = RHO(1,2)^2;
    end

        % R² ajustado
        R2a=1-(erro/(n-k-1))/(SST/(n-1));

        %Log R2 ajustado
        logR2a=abs(log(R2a));

        %R2 modificado (sugestão Jorge - Dissertação Lisandra)
        RR=-log(1/(1-(1-(erro/SST))));

        [R2a RR];

        %Teste de hipótese para cada parâmetro estimado
         s=sqrt(erro/(n-k-1));

         alfa=0.05; % nível de confiança desejado
```

```
          talfa=tinv(1-alfa/2,n-k-1); % valores de t de referencia para o
nível de confiança desejado

          t=[];
          for i=1:nw+1
              t(i)=(par(i)-bj0)/(s*sqrt(c(i,i)));
              if abs(t(i))<talfa
                  t(i)=0;
              end
          end

         % Tamanho do intervalo de confianca

         for i=1:nw+1
             intervalo(i)=2*talfa*s*sqrt(c(i,i));
         end
         intervalo;

      SSR=SST-erro;

end

end
```